Requirements for Maintaining Web Access for Hearing-Impaired Individuals*

DANIEL M. BERRY dberry@uwaterloo.ca
School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1 Canada

Abstract. The current textual and graphical interfaces to computing, including the Web, is a dream come true for the hearing impaired. However, improved technology for voice and audio interface threaten to end this dream. Requirements are identified for continued access to computing for the hearing impaired. Consideration is given also to improving access to the sight impaired.

Keywords: access, closed captioning, e-mail, fax, hearing impaired, lipreading, lipsynching, movies, sight impaired, talking head, telephone, textual and graphical interfaces, TTY, TV, video telephone, voice and audio interfaces, voice synthesis

1. Introduction

A paper published in a previous issue of this journal “Quality in Web Design for Visually Impaired Users” by Margaret Ross (Ross, 2002) describes the specifics of insuring and measuring the quality of Web sites from the perspective of the sight-impaired (SI) individual. These SI individuals include the blind, the partially sighted, the elderly with deteriorating eyesight, the color blind, and the dyslexics. The paper discusses requirements for, legal issues for, and testing for accessibility of Web sites to the SI individual. It also surveys specific Web sites that have been tested for this accessibility.

The purpose of the present paper is to discuss accessibility of the Internet, e-mail, Web sites, and other communication devices from the perspective of the hearing-impaired (HI) individual.

2. Background

I am hearing impaired (HI) from birth and understand spoken language mostly by reading lips. Thus, I have always had problems using a telephone, which shows no lips. I have always been more comfortable with written communication. I have been using computers since 1965 and have been using the ARPA Net and later the Internet for communication

* This paper is an expansion of a paper published by the same author in the *Proceedings of the Third International Workshop on Web Site Evaluation (WSE’01)* (Berry, 2001).

since 1979. Computers, up to now, have been a boon to me, and for that matter to the rest of the HI world. In particular, they allow me to communicate with nearly all of my circle of acquaintances, a large fraction of which are in the computer business, by textual and graphical means, i.e., by e-mail, by Web page interaction, etc. For the few acquaintances that do not have e-mail, fax usually is available.

More recently, telephones have become even more difficult to use. I feel that the equipment available today is of markedly lower quality than the equipment we used to rent from Western Electric, and there is more distortion when the sound is amplified.[1] In addition, the increased use of answering machines, voice mail, and voice-directed menu selection[2] have taken away the possibility of my asking the person on the other end of a call if I understood her[3] or of my requesting her to repeat what she just said. In essence, I have become disenfranchised from the telephone, so much so that I do not give out my telephone number any more. This disenfranchisement was not so bad, since it was always difficult to use the telephone, and in any case, computers provided an alternative communication means that has become almost as universal as the telephone, at least among those with whom I want and need to communicate. Quite naturally, I have a vested interest in keeping things the way they are.

The current work (Leavitt, 2003; Marcus, 2003; Wang, 2003) being done to build voice interfaces to computers worries me. I see that speech recognition algorithms are achieving more than 95% accuracy (Leavitt, 2003). Applications that depend on accurate speech recognition are being built and deployed, including to drive e-commerce applications (Fainchtein, 2002) and, ironically, to drive software that helps the HI individual by showing lipreadable lips mouthing out the speech that is being recognized (SpeechView, 2003). Thus, I feel that computers may be going the way of telephones towards my disenfranchisement. I watch Star Trek, taking place some 250 years in the future and see people interacting with the shipboard computer by talking with it. I personally would prefer that computers stay with entirely textual and graphical interfaces (TGIs). Clearly, I cannot stop the deployment of voice and audio interfaces, i.e., sound interfaces. Also, strictly TGIs are a problem for sight-impaired (SI) people, who prefer sound interfaces. Therefore, by this paper, I attempt to prevent my total disenfranchisement by recommending changes to the future directions that will make it possible for me, and the rest of the HI world, to continue to work with computers and to use computers for communication.

I feel that my disenfranchisement from the telephone happened partially because people like me did not complain enough, probably because an alternative was becoming more usable at the same time. Thus, I believe that it is necessary for HI individuals to take active steps to prevent their disenfranchisement from the computer, the Internet, and the Web, that is, to *maintain* Web access for the HI individual.

The problem is not just mine. According to 1990 and 1991 surveys by the National Center for Health Statistics, approximately 8.6% of the U.S. population three years and older have hearing problems, and that among these, 2.75% are profoundly deaf (Signtel Inc., 2003a).

To provide justification for the proposals, Section 3 builds a model of what an HI individual can and cannot do and why. To clarify this model, the appendix gives details about my own hearing as one point in the model's space; while I am unique and atypical in many ways, I share many attributes, problems, limitations, solutions, needs, and hopes with all HI people. Section 4 observes that the HI people and the SI people have conflicting requirements. The proposals are presented in Section 5. Section 6 describes other work towards the same goal.

3. Abilities and classifications of HI persons

According to traditional audiology, understanding speech requires being able to hear with no more than a 75 decibel (db) loss in the range of 500–, 2000 Hertz (Hz). Figure 1 shows my audiogram with this requirement represented as a rectangle bounded by a dotted line.

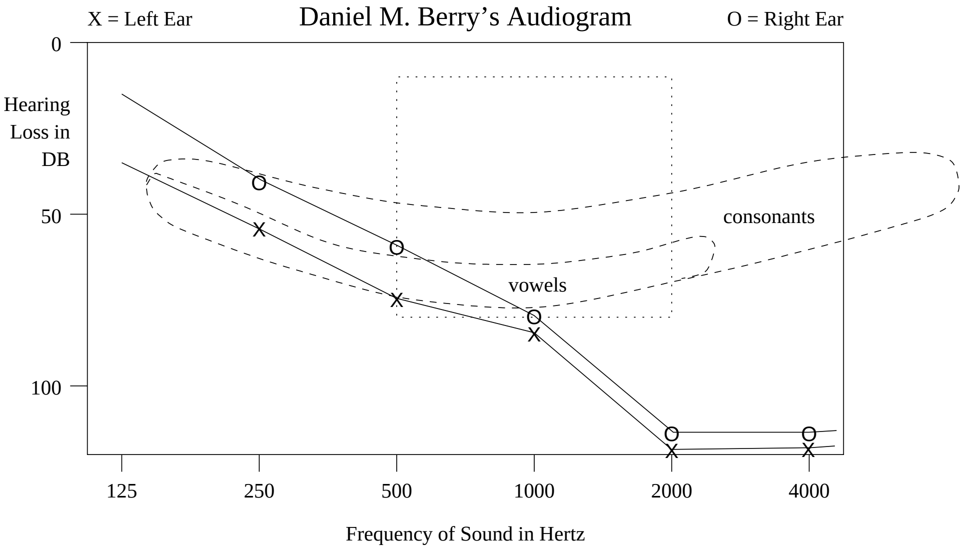

Figure 1. Audiogram

An audiogram shows two plots, one for each ear. The plot for an ear shows for each frequency, the hearing loss of the ear at the frequency. The loss of an ear at a frequency is measured by determining the minimum volume required for the ear to hear a tone of the frequency. The more of the speech-understanding rectangle that lies below the plots for an ear, the more that the ear can help understand human speech. More recently, the regions required for hearing vowels and consonants have been mapped. They give a more accurate way to determine whether or not a person can understand speech and to identify which part of it he does. The more of these regions that lie below the plots for an ear, the more that that ear can help understand the vowels and consonants, respectively. Note that the vowel region is entirely contained within the consonant region, since some consonants, e.g. “m”, are not just explosions and have a voice component, as do all vowels. Note also that according to the speech-understanding rectangle, I appear to understand much less than I know I do; the vowel and consonant regions model my understanding more accurately.

There are several independent ways to classify an HI person, by

1. severity of his hearing loss,
2. length of time he has had the hearing loss, and
3. kind of input he requires in place of pure voice.

This classification is at best a guide for an initial guess as to what the HI person is able to do. Many individuals do not fit exactly into the classifications, and the capabilities of many individuals differ from what I claim is typical for persons in each classification.[4] Nevertheless, the reader should gain an appreciation for what is possible and what is needed in Web interfaces to accommodate the HI individual.

3.1. Severity-of-loss classification

When classifying an HI person according to severity of hearing loss, there are three groups:

1. A person in the first group has less than a 50db loss in all frequencies; that is, he has some usable hearing in all frequencies.
2. A person in the second group has greater than 100 db loss in all frequencies; that is, he is considered totally deaf.
3. A person in the third group is in neither the first nor the second group. He has usable hearing in some ranges of frequencies and is totally deaf in other ranges of frequencies.

I happen to be in the third group.

Typically, a person in the first group speaks fairly well and wears a hearing aid that amplifies all frequencies. With such an aid, the person functions about as well has a non-HI person. Typically, a person in the second group only signs and does not wear an aid, which is actually quite useless for his hearing. However, very rarely, a person in the second group has been trained to make use of the very tiny residual hearing he does have with the help of a hearing aid and with or without lipreading. In the third group, a smaller majority only sign. Less rarely than in the second group, a person of the third group uses the hearing he does have with the help of an aid and with or without lipreading. The reason that most of the second and third group sign is that for historical and traditional reasons, most of them are sent to schools for the deaf in which they learn signing and are not taught to make use of the hearing they do have.

A person in the first group may be functionally not HI, especially if he is using a good hearing aid.

3.2. Length-of-time-of-loss classification

When classifying an HI person by the length of time he has had the hearing loss, there are two groups:

1. A person in the first group has loss his hearing since before he could talk, i.e., during birth or infancy.
2. A person in the second group has loss his hearing after he learned to talk, i.e., during youth or adulthood.

I happen to be in the first group.

This classification is fuzzier than most, but the key questions to ask about the instant in which the person lost his hearing are:

1. Has he already learned to speak normally and thus can continue to make sounds correctly even though he can no longer hear what he is supposed to be imitating?
2. Does he already know what speech normally sounds like and thus knows what he is missing?

Someone in the first group answers “no” to both questions and someone in the second group answers “yes” to both questions. It is hard to imagine someone giving a different answer to the two questions.

The typical person in the first group behaves as predicted according to the severity-of-loss classification. The typical person in the second group speaks quite well but has difficulty understanding speech because he has had to relearn hearing or to learn lipreading or signing at an age in which acquisition of a new language or even a new form of input for a familiar language is very difficult. This difficulty seems to be independent of the severity of the loss and has more to do with age and the ability to learn new languages.

A person in the second group may be functionally not HI, especially if his hearing loss is not severe or he is wearing a good hearing aid.

3.3. Kind-of-input classification

When classifying an HI person according to the input he requires, there are three groups:

1. A person in the first group requires signing.
2. A person in the second group uses a combination of residual hearing and lipreading to understand speech as it is spoken.

3. A person in the third group uses only residual hearing.

I happen to be in the second group.

A person in the first group has never really learned to handle arbitrary speech, and even a hearing aid does not make it possible for him to understand speech without use of the alternative input medium such as signing or text. A person in the second group generally wears an aid. Usually, he also signs, particularly if he has a lot of acquaintances that are also HI. A person in the third group typically has a mild loss that is uniform over the spectrum. He can generally get by in the hearing world if he is assisted by a hearing aid that corrects the loss.

A person in the third group may be functionally not HI, especially if his hearing loss is not severe or he is wearing a good hearing aid.

Many HI signers cannot read lips at all. Among those that do read lips, many do so poorly and could not rely on lipreading for total and accurate input. The typical HI signer is communicating only by signing. He has very poor speech, which is very difficult for a non-HI person to understand without getting used to it. He interacts only with other signers, whether they be HI or non-HI that have learned signing, e.g., his non-HI close relatives and friends. He is not able to hear on the telephone and uses TTY[5] in place of the telephone to communicate with his HI acquaintances, with relatives and close friends who have TTY units and with organizations offering TTY lines. He reads and writes and can use computers, e-mail, and fax. He requires captions or subtitles on TV shows or movies. These signers are the largest group of HI individuals that have to be accommodated on the Web.

3.4. Summary

However different the abilities of HI persons are, for any given HI person, unless he is functionally not HI, the basic fact is that he cannot depend on auditory input, and such auditory input must be replaced by or augmented by visual input.

4. The HI people and the SI people

It should be clear what is good for the HI person is not good for the SI person and vice versa. Right now the Web is perfect for the HI person and not so good for the SI person. However, the SI people are complaining, and legitimately. As a result of the complaints of the SI people, R&D exists that is directed towards enfranchising the SI. That enfranchisement can easily come at the expense of the HI people, possibly even disenfranchising the HI people. There is no need for the HI people and SI people to be competing. Therefore, this paper is recommending ways that prevent the disenfranchisement of the HI people without impeding progress to enfranchise the SI people.

My recommendations are valid for all HI people, providing, when possible, also for those who do have some auditory input and oral output. I take into account also the SI people who cannot use text and pictures directly, but can use text converted to voice or textures, e.g., in the form of Braille.

As a basis of my recommendations on behalf of the SI people, I am using the experiences of a blind student that took one of my courses recently. He had difficulty with the electronic copies of my slides and the course Web page, particularly when these involved pictures and diagrams. He was able to read the text of these through a device with earphones that could read ASCII or scanned text and pronounce what it read. He can also read Braille.

5. Recommendations for sound-based human–computer interfaces

At the highest level, my recommendations are:

1. When the computer speaks to the user, it should do so both by sound and text or pictures, and that the sound and text be synchronized to minimize the cognitive interference that happens when captions are shifted too far from the video that they caption. An added bonus would be to have a visible talking head mouthing out the sound, to allow those who read lips to do so rather than to have to read the text.
2. When the computer is to accept input from the user, it should accept both voice and textual input. Many HI people are not able to speak well or consistently, and many SI people find that typing is difficult.

5.1. Output from the computer

As mentioned, when the computer outputs to the user, it should be both in sound and text or pictures. The specifics of this recommendation depends on which medium is the original source and thus, which other media has to be generated from the source.

5.1.1. Source is text If the source is text, then the sound can be generated by a voice synthesizer that is operating on the text, such as what my blind student had to read ASCII files. Providing a talking lipreadable head synchronized with the generated sound would require use of the technology of lipsynching (Martin, 2003; Comet, 2003; Third Wish Software, 2003). Figure 2 shows snapshots taken at key points[6] during the animation, produced by Michael B. Comet (Comet, 2003), of visemes[7] for the phonemes[8] of the English language. The reason several phonemes share the same viseme is that all phonemes in the set under a particular viseme appear the same on the lips when spoken. Lipsynching allows animation of faces having lipreadable lips synchronized with sound. However, the talking lipreadable head is not essential if the source is already text.

If the source is text in a phonetic alphabet designed to make voice synthesis easier, then this phonetic text should be displayed. HI people who watch real-time close captioning are used to dealing with incorrect spellings that yield correct pronunciations. It would take such a person a short time to get used to reading the phonetic alphabet.

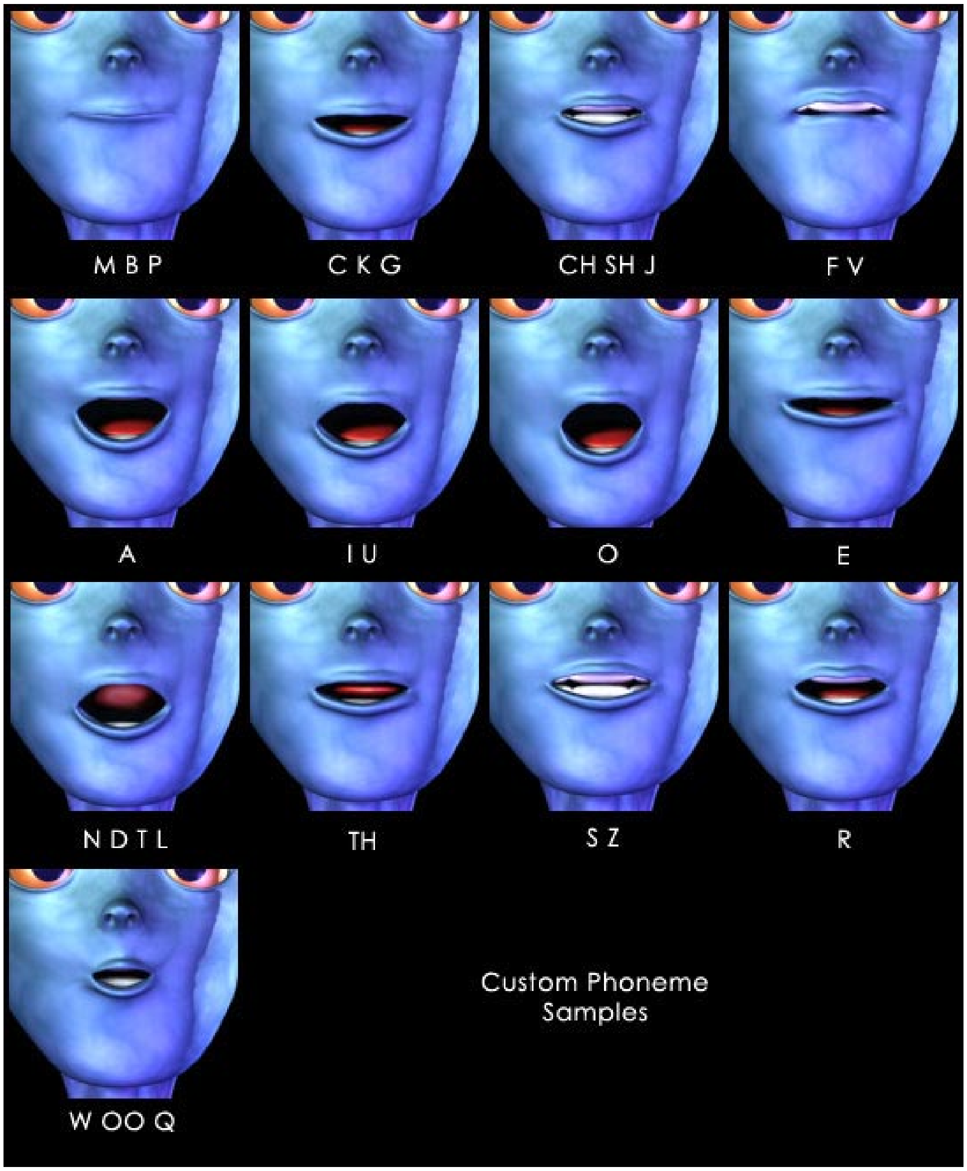

Figure 2. Animated Lipreadable Lips by Michael B. Comet

5.1.2. Source is a real person's voice If the source is the voice of a real person, then a video of that person can be made as he is being recorded. This video would provide the lipreadable talking head. In this case, captioning is necessary to augment the video and sound. If the person is reading a script, then the script can be displayed, as is done with closed captioned pre-recorded TV shows and movies. The captions should be synchronized with the sound.

For alive video, presenting the text requires real-time captioning by a person with the skills of a court-room stenographer, as is done for closed captioning of alive television, e.g., the news or sporting events. Perhaps in the future, automatic voice and speech recognition will have advanced to the stage that this software can provide captions in real time. It is not clear that the 95% (Leavitt, 2003) or better accuracy rate of the current speech recognition algorithms is good enough for sufficiently accurate captioning. However, apparently that accuracy is good enough to produce acceptable animated lipreadable lips (SpeechView, 2003). Since several phonemes share the same viseme, it might very well be that the mistakes made by the speech recognizer map an incorrectly recognized phoneme to a phoneme that happens to share the same viseme.

For previously recorded video such as of movies and pre-recorded TV shows, captions, if available, should be shown. If captions are not already in the video, then they need to be added. In any case, the captions should be synchronized with the sound.

5.2. Input from user

The computer should be prepared to accept input by a variety of means without the user having to announce beforehand the preferred form of input. That is, at any time, the computer accepts and interprets input from whatever medium the input comes.

The means of input that can be accepted are

1. voice, powered by voice recognition technology such as IBM's ViaVoice (IBM, 2003),
2. keyboard, typing a direct response, and
3. mouse, clicking on buttons or menu entries or making gestures.

Nowadays, a copy of personal voice recognition software residing on a computer can be trained to understand the limited number of designated users of the computer on which the copy resides. The accuracy of recognition is high enough that this user can almost completely dispense with the keyboard and pointer (mouse). However, if the user has difficulty speaking clearly and consistently, as do many HI people, voice input may not work reliably, and the other means of input will be needed.

5.3. Summary

Looking back over the recommendations, it appears that a textual interface is the key.

The HI individual who is not SI can function with text. Moreover, from text, one can synthesize other representations, such as large letters, Braille, and voice, that can help the SI individual. While to generate other media from text is straightforward, generating text from other media is not even algorithmic in many cases. We still cannot generate text reliably from voice. Thus text is the simplest basis representation.

6. Other work

Among the other work are work done by the W3C, some international organizations, private companies, and governmental organizations.

6.1. W3C guidelines

Just as the conference version (Berry, 2001) this paper was accepted for publication, ACM's *Interactions* published W3C's "Web Content Accessibility Guidelines 1.0", dated 5 May 1999 (Chisholm, 2001). The report is noteworthy to me because it goes to the heart of my own recommendation. The report and my own recommendations amount to independent confirmations of the same ideas.

The W3C report's main recommendation is that text should always be available for any artifact. "The guidelines do not suggest avoiding images as a way to improve accessibility. Instead, they explain that providing a text equivalent of the image will make it accessible.... Text content can be presented to the user as synthesized speech, braille [sic], and visually-displayed text. Each of these three mechanisms uses a different sense—ears for synthesized speech, tactile for braille, and eyes for visually-displayed text—making the information accessible to groups representing a variety of sensory and other disabilities.... While Web content developers must provide text equivalents for images and other multimedia content, it is the responsibility of user agents (e.g., browsers and assistive technologies such as screen readers, braille displays, etc.) to present the information to the user."

If an artifact is not readily textual, a functionally equivalent textual representation should be available. That is, if the artifact is a digitized photograph of a house,

1. and the purpose of the picture is to show the viewer a pleasant scene containing a house, the alternative text for the picture should be something like "photograph of a pleasant scene containing a house"
2. and the purpose of the picture is to be an icon for transferring to the home sales department, the alternative text for the picture should be something like "transfer to the home sales department"
3. and the purpose of the picture is to sell the specific house pictured, the alternative text for the picture should be a detailed description of the house, for example, "picture of newly painted wood-frame house with three-bedrooms, two and a half bathrooms, large kitchen, two-car garage...."

The reader is urged to consult the published report or the Web page for more details.

The UK Center for Applied Special Technology (CAST) has developed an automatic test of Web site accessibility based on the W3C's Web Content Accessibility Guidelines (Ross, 2002). The test, known as the "Bobby Test", is available at the Bobby site (Watchfire Corporation, 2003). When the test is applied to a URL, it automatically inspects the site referenced by the URL for accessibility according to the guidelines. The test creates a copy of the referenced site's pages, marked with Bobby (British Police) hats and question marks. A Bobby hat with a wheelchair denotes a spot of very poor accessibility, and a question mark denotes a spot that the automated test cannot check and that must be checked by a person. A report is given of all the problem spots at all three levels of severity. Finally, the site is given a grade, ranging from a low of "Nonconforming“ to "Conformance Level A" through "Conformance Level AAA" to a high of "Conformance to U.S. Section 508 Final Rule". A conforming site is allowed to put the Bobby Logo on its pages.

6.2. International organizations

There are other organizations dealing with Internet access for disabled people, including

1. ICDRI, the International Center for Disability Resources on the Internet (ICDRI, 2003), and
2. EASI, Equal Access to Software and Information, located at Rochester Institute of Technology (EASI, 2003).

6.3. Private companies

There is a company, Signtel (Signtel Inc., 2003b), that builds assistive technology for the HI for use by on-line organizations. The company has developed some of the technology that is needed to implement the suggestions of Section 5. In particular, it has developed software to map

- from speech to text,
- from text to sign language,
- from text to speech, and
- from text to moving lips

and to do so synchronously, so that the various media can be used to complement each other.

There is yet another company, SpeechView (SpeechView, 2003), that has developed LipC, software that resides on a computer connected to a telephone whose own handset has been disabled or disconnected so that the computer can serve as both the input and the output of the telephone. The software listens to the sound being transmitted to the telephone and computes animated lipreadable lips, i.e., visemes, for the phonemes it finds in the sound. The software outputs the sound on the computer's speakers and displays the animated lips on the computer's screen. The software delays the sound output a bit to allow the algorithm time to find the phonemes, to compute the visemes, and to display the visemes synchronized with their phonemes.

SpeechView has a creative approach to dealing with the fact that several phonemes share the same viseme. SpeechView's philosophy is to give LipC users a minimal set of self-explanatory signs that allow differentiating phonemes that share a viseme, being careful not to overload the user with information that makes interpretation less automatic. LipC has three signs:

1. Change the color of the throat when the phoneme is a voiced consonant.
2. Change the color of the nose when the phoneme is a nasal consonant.
3. Place colored circles on the cheeks when the phoneme is an explosive[9] consonant and there is another, necessarily non-explosive phoneme with the same viseme that has no distinguishing signs.

To make sure that the user sees the additional signs, LipC prolongs display of the additional signs a bit into the viseme for the following vowel phoneme. For example, the difference between "big", "mig", and "pig" is that

1. "big" shows a colored throat on the "b" and part of the "i";
2. "mig" shows a colored throat and a colored nose on the "m" and part of the "i";
3. "pig" has no additional sign at all even though it is explosive, because the other visemes of the same phoneme show additional signs.

Thus, the lipreadable lips presented by LipC are better for the lipreading HI individual than are natural human lips, because a real person's throat, nose, and cheeks do not change color as she is speaking!

The company is Israeli and has chosen to focus on the Hebrew HI market first. Consequently, at this time, the SpeechView software is available only for Hebrew. The company plans to release an American English version in 2004. Moreover, the company has chosen to focus first on land-line telephones rather than cellular telephones. The interface between a land-line telephone and a computer over a wire is simpler than between a cellular telephone and a computer. Also, traditional telephones with their better quality sound are more common among HI individuals than are cellular telephones with their poorer quality sound.

6.2. Governmental organizations

In an effort to comply with the requirements of the UK Disability Discrimination Act, the UK Post Office began the TESSA (TExt and Signing Support Assistant) Project to build special user interfaces for deaf customers to allow them to purchase postal services and licenses just as anyone else does (SYS Consulting Ltd, 2003; Lincoln, 2001; ViSiCAST, 2003; Boyd, 2002). The deaf customer interacts by talking or typing with a human postal agent, who is competent to make intelligent decisions that would be beyond the scope of software. To talk with the deaf customer, the agent engages a voice recognition system that generates animated signing on the screen of the deaf customer.

7. Conclusions

In this paper, I have given some recommendations of things that will help keep computers accessible to the HI population while affording more opportunity for the SI population to use computers. I have described the various kinds of hearing impairment, including my own, to motivate and explain my recommendations.

The recommendations do not require any new technology or research. They required only understanding the problem and the solutions, being aware of opportunities to solve the problem, and being careful to apply the recommendations as Web page structure and content are planned.

Acknowledgments

I thank Eran Aharonson, Nitza Avni, Helen Beebe, David Brown, Mike Burks, Meir Burstin, Vint Cerf, Michael Comet, Igor Fainchtein, Emil Froeschels, Craig Gotsman, Antoinette Goffredo, Naomi Heller, Nachshon Margaliot, and Mike Melkanoff for their teaching, discussions, or input. I thank Jim Bieman, Margaret Ross, and the anonymous referees of earlier drafts of this paper for detailed criticisms that allowed me to improve the presentation considerably. I was supported in part by NSERC grant NSERC-RGPIN227055-00.

Appendix. My hearing, speech, and communication

This is a personally motivated paper. Therefore, a little background about me is useful. Also, I am a concrete example of the general HI person described in Section 3.

A.1. My hearing

I am HI since birth. I do not sign, but I do read lips. I read lips well enough that people forget that I do not hear very well and that I cannot understand any sound device that does not allow me to see the speaker’s lips, such as the telephone. Notice that in the author’s address information in this paper, I explicitly list no telephone number; instead I direct people to send me faxes or e-mail.

I hear a little, with a 50 db loss, at frequencies below 500 Hz. Thus, I can hear vowels and sounded consonants such as “m” and “b”. I am essentially totally deaf, with a 110 db loss, at frequencies above 1000 Hz. Thus, I cannot hear non-sounded consonants such as “s” and “p”. My audiogram, shown in Figure 1, shows that my hearing misses most of the rectangular region considered essential for understanding speech. Clearly, I cannot follow normal speech because so many of the sounds are missing. That is, with the sound that I hear, the language is too ambiguous. To me, with sound alone, each of “cam”, “fam”, “ham”, “kam”, “pam”, “qam”, “ram”, “sam”, “tam”, “wam”, and “xam” sounds like “am”.

I wear a hearing aid to help me make better use of the little hearing I do have. A hearing aid that amplifies every frequency would be counter productive since it would amplify beyond comfort that which I can hear without it, and it would amplify low-frequency background noise to the point of distraction. Therefore, I wear a special, prescription hearing aid. The amount of amplification at any frequency below 1000 Hz decreases with the frequency. Since I have no hearing at all above 1000 Hz, it does nothing to those frequencies. Also since my hearing decreases with increasing frequency, it shifts frequencies below 1000 Hz a bit lower, although not enough to cause me to lose the ability to distinguish voice tones sufficiently to read emotions.

The hearing aid has also a telephone coil. This coil is actually a radio receiver that picks up the radio waves generated by the electromagnetic oscillator in the good handset speakers. By picking up radio waves, the sound I hear has not suffered any distortion by transmission through the air; the sound is generated inside the hearing aid. Unfortunately, there are handsets that do not work with the telephone coil; they use carbon oscillators that do not generate electromagnetic waves in addition to the sound waves. Carbon oscillators are found on the cheaper handsets and on many cellular telephones.

A.2. My lipreading

I read lips to fill in on the missing sounds. I learned to read lips the same way that most people learn to understand spoken language. As a toddler, I began to notice patterns of lip movements, i.e. visemes, and the phonemes that I heard that were highly correlated with meaning, just as the average person notices patterns of phonemes that are highly correlated with meaning. To the average person, the sound patterns are sufficiently unambiguous, that lip movements are not needed to disambiguate. In my case, with the addition of lipreading, all of the words above that sound like “am” are distinguishable from “am” and each other.

Lipreading itself is not unambiguous. It is a lot less ambiguous that the portion of speech that I hear, but is a bit more ambiguous than speech for the hearing person. Specifically, some different phonemes share the same viseme. For example “m”, “b”, and “p” appear the same and so do “d” and “t”. I said that this is a slight ambiguity, because even hearing people deal with this sort of ambiguity; “k” and “c” followed by “a”, “o”, or “u” have the same phoneme, but people distinguish words containing them by context. In my case, I am able to hear “m” and “b”, but cannot hear “p”. So if the viseme appears to be one of them, and I cannot hear the phoneme, the phoneme must be a “p”. This decision is carried out entirely subconsciously, just as distinguishing the different meanings of a homonym. Therefore, I need the sounds I hear to disambiguate the phonemes with the same viseme. Thus, I cannot read lips when there is no voice or in noisy rooms, because I am lacking some important disambiguating information.

This need of voice to disambiguate phonemes with the same viseme is quite personal and is language dependent. Other HI people with less hearing do not hear even “m” and “b”, but they have learned as effortly as the hearing person learns to distinguish homonyms, to use language knowledge and context to distinguish between “m”, “b”, and “p”. The lips for “micro” are definitely saying “micro” because “bicro” and “picro” and not words, and knowledge of the context can tell the listener whether the word is “Mom”, “Bob”, “Pop”, “mop”, “mob”, “bomb”, or “pomp” after language knowledge has eliminated the other combinations. In Hebrew, there is a group of eight letters that have the same viseme and have phonemes that are outside of my hearing range. So I have trouble with Hebrew. There are native Hebrew lipreaders. Thus, the ambiguity introduced by these eight letters must be manageable for the native speaker.

I am able to read lips from the side, and the lips of a non-native speaker of English speaking with a heavy accent seems not to faze me. However, I do have problems reading lips and understanding native speakers of Australian English, known as Strine (spelled “Australian”), and of the Scottish brogue.

A.3. My speech

My native, natural speech is a reflection of what I hear and lipread, just as the hearing person’s natural speech is a reflection of what he or she hears. I do not hear the letter “s” at all and recognize it only by its lip and teeth configuration. Thus, in my natural speech, when I intend to say “s”, my lips and teeth go to the right places, but there is no sound. My pronunciation of “Sam” is “am” preceded by my lips and teeth being right for “s” for the right amount of time, but with no sound. Later, as a teenager, I was trained to make sounds I cannot hear. However, since I cannot hear them, I cannot be sure that I make them correctly or even at all. I am quite sure that I sometimes do not.

A.4. My communication

My hearing, lipreading, and speech contribute to a particular pattern of communication in which I do certain things to ensure understanding of speech and in which I avoid things I cannot do.

A.4.1. My conversations In order for me to listen to or converse with someone, I need to position myself so that I can both hear her voice and see her lips. Lectures, when I can sit close enough to the speaker, and one-on-one conversations are easiest. When the number of people in a conversation is more than three and the conversation moves randomly around the group, I get lost. By the time I have found the person who is speaking to read her lips, I have missed the first sentence or so. I end up missing portions of the conversation that are essential for following the conversation. Hence, I shy away from large groups and parties.

When I follow the conversation by lipreading, I interact well enough that people forget that I am HI. I sometimes have to remind people to face me or to not cover their lips.

A.4.2. Other languages I read, write, and speak several languages besides English, namely French, German, Hebrew, Portuguese, and Spanish. However, I am not able to understand any of them spoken. I speak them well enough that people answer me in the language I speak. Therefore, it is dangerous for me to speak these languages, because I quickly get responses that lose me. The reason I cannot understand these spoken is that I cannot read lips in them. I have tried to learn to read lips in Hebrew by taking lessons and living in a Hebrew-speaking environment, in Israel, but even after three years of lessons and eleven years living in Israel, I was not able to break loose from the low plateau on which I was stuck. I later learned from a speech therapist in Los Angeles specializing in lipreading that learning to read lips in anything but one's native language after the age of 5 is virtually impossible.

A.4.3. signing I do not sign. Therefore, a signing interpreter is of no use to me. As a side effect of not signing, I have very few HI acquaintances.[10]

A.4.4. Telephone use I generally cannot understand what the person on the other end of a telephone conversation is saying because I cannot see her lips. If, however, I am controlling the telephone conversation and have constrained the subject or am asking yes-or-no questions, then I can follow what the other person is saying. In the first case, the possible answers are constrained enough that I can often hear enough of the words that I can tell which of the possible answers it might be. Then I can ask yes-or-no questions to confirm that I have heard them correctly. My hearing is good enough that I am able to distinguish “Yes” from “No” without reading lips; the vowels, which I can hear are quite distinctive. I have learned to structure many conversation so that I can get all the information I need by asking strategic yes-or-no questions. While numbers are difficult to distinguish, I can ask the other person to count up to each digit.

Apart from these highly constrained situations, I cannot understand the other person, particularly if I am not expecting such a call and have no idea what the call might be about. I am often not even able to understand the name of the person who is calling.

Therefore, I generally do not answer my telephone. I use the telephone mostly only for incoming and outgoing faxes and outgoing telephone calls that I can control. I have caller ID allowing me to see who is calling if she has not disabled my seeing that information. I make an exception and answer an incoming call when I can identify the caller and it is someone that I know well and can thus guess what the conversation might be about. On my home telephone, so that people do not assume that I am not at home for long periods, I have a recording saying that even if I am at home, I do not answer the telephone and to please send a fax to the same number.[11]

I cannot use a cellular telephone or remote handset, even when I am controlling the call. Unfortunately most such equipment does not have the required volume or if it does, it distorts too much at the high volumes so I cannot even understand “Yes” and “No”. Many of them have only carbon oscillators that do not broadcast to the telephone coil in my hearing aid. In fact, the *only* telephones I can use are the old Western Electric 500 standard telephones. The handsets have such good undistorted sound that I can hear what I do hear even without amplification so long as I am using the telephone coil on my hearing aid. As mentioned in Note 2, it seems that because these telephones were built for rental and AT&T had to replace them free of charge if there were any damage, they were built so well and so far beyond the minimum threshold that even with maximum amplification they are not near the equipments limits. Since the so-called liberation of the telephone services about 20 years ago when we had to start buying our equipment, the quality has gone down hill. Fortunately for me, these old telephones are indestructible. So, I have saved them and continue to use them. At the same time, the rest of my family has had to replace many broken telephones bought from a variety of telephone manufacturers.

If I am in a situation in which I need to make a telephone call and I do not have the right equipment and I cannot be in control of the conversation, I ask someone else to be my ear, even when I am asking for a date!

A.4.5. Recording and IVR The bane of my life are recorded messages, left for me in hotel rooms or played at numbers that I have called. Even if the subject is controlled, I have no way to confirm with the recording that I have heard it correctly. Moreover, the quality of the recoded voice is never as good as a real voice. What I hate the most is Interactive Voice Response (IVR), namely the automatic, recording-directed menu selection regime that is so common these days when one calls an institution. I am referring to these recordings that say “Welcome to XXX. If you want to deal with AAA, press 1 now. If you want to deal with BBB, press 2 now, ... and if you wish to speak to a customer service representative, please stay on the line.”

Not only do I have all the problems of understanding the recording and not being able to ask if I understood correctly, but also if I take a chance and hit the wrong key, I tend to get into a state from which I cannot escape, because I do not always understand what is being said to me. Moreover, it seems like I am put on hold forever when I choose to stay on the line to speak to a human being. I am not even sure that there *is* a human being, because I cannot be sure that the recording did say, “Please stay on the line to speak to a customer service representative.”

A.4.6. E-mail and fax Thus, for telephone-like communication with others, I use mainly e-mail and fax. Most of my acquaintances are computer people or their relatives.

So, most people I know have e-mail and have had it for years. With the popularity of the Internet these days, more and more of my other acquaintances have e-mail. Nowadays, when I meet a new acquaintance, female or male, I ask for an e-mail address instead of a telephone number, and my request is usually granted. These days, the few acquaintances that do not have e-mail are businesses that have not computerized. Almost all of these have fax. So it is very rare indeed that I have to use the telephone.

However, today the widespread use of e-mail is threatened by the much maligned, ever-growing epidemic of spam. A majority of the messages in a typical person's mail box is spam. A person often deletes or ignores non-spam messages in his or her zeal to quickly get to the non-spam messages, thus increasing the chances that e-mail from me will be ignored. I hope that something will be done to eliminate spam before people get so turned off from using e-mail that they stop entirely.

A.4.7. TTY Many HI people use TTY units with the telephone in order to be able to communicate with others via a telephone with text. Two people with TTY units at the opposite ends of a call connection type to each other in real time, much as with the UNIX talk command, except that the screen is not split into send and receive windows. The sent and received text are interleaved. Hence, the conversers have to set up a protocol to prevent the two from talking, i.e., typing, at once.

Many institutions provide TTY numbers and operators to allow HI people to interact with them. A TTY unit consists of basically an old fashioned hard copy (key and ribbon) terminal together with a 150–300 baud modem operating with an ancient 5-bit character code called Baudot. Baudot was the code used before ASCII and it was adopted for TTY so that the HI community could get discarded equipment cheaply as the rest of the world adopted ASCII.[12] I do not use TTY because no one I communicate with has a unit. There are less than a handful of HI people in my circle of acquaintances, perhaps because I do not sign. Each of these HI people happens to use e-mail like I do.

A.4.8. Watching TV or movies I cannot watch TV or movies by lipreading alone, since not always is the speaking person facing the camera. Some TV shows and movies have narration from off screen. I watch only TV shows and movies that are subtitled or that have closed captioning. I do not go to theaters except for subtitled movies. I wait until movies appear on video tape or DVD, and I boycott movies and producers that make non-captioned videos.

When I go to a place in which French, German, Portuguese, or Spanish is spoken, and I am able to follow English speaking movies that are subtitled in these languages. I can read these languages fast enough. While I can read Hebrew, because of its non-Latin alphabet, I cannot read it fast enough to be able to follow Hebrew-subtitled English-speaking movies; each subtitle line disappears before I have finished reading it.

A.4.9. Video conferencing Quite clearly, it is impossible for me to participate in meetings conducted with a conference call or with a speaker telephone. Assuming that a face-to-face meeting is not possible, then only video conferencing has a possibility of working for me, as the possibility exists to read lips. I was once in a meeting in which the video was transmitted over an expensive high-speed dedicated line, and the refreshing of the video was at the standard TV rate, which is high enough that it was possible to read lips. So long as the speaker faced the camera, I fared well. However, most of the time, the video conferencing is done over a cheaper standard telephone call connection or over the Internet, and the update of the picture is not frequent enough to show smooth lip movement. Consequently it is impossible to read lips. As the bandwidth of telephone lines increases, this problem will solve itself.

A.5. Technology that I would love to have

I am waiting for the day when video telephone use is widespread enough that everyone with whom I interact has one. Then I would get video telephone and would be able to lipread over the telephone. There are video telephones available now. Even ignoring the fact that not enough people have them, there is a problem inhibiting their use for lipreading. The current bandwidth available for video telephones allows the video to be updated less frequently than is required for alive action. The consequence is that the picture is updated infrequently enough that the video is really a sequence of disjoint stills rather than a continual stream in which the lips appear to move. If I understand correctly, the designers of the video telephone had a choice as what to allow to degrade, the video or the audio. Based on the needs of most of the population, which hears well enough, it was decided that audio quality is more critical and that to see the person to which one is talking and to see where that person is, stop-motion video is sufficient. Stop-motion video might even be enough to read body language. However, for me and other HI people, the opposite choice should be made. That is, it would be preferable to me and them that the audio degrade to preserve video quality. I could probably get enough of the voice to disambiguate lipreading from degraded audio.

Since each user is different, the best would be to give a means for the user to choose what to degrade and by how much, perhaps with a slider stretching from 100% video quality to 100% audio quality.

A different approach is that of SpeechView (SpeechView, 2003). If one cannot get enough bandwidth to send a sufficient number of high quality images for alive-motion video, perhaps the images of lipreadable lips can be generated locally based on a sufficiently accurate analysis of the sound that is being sent. As mentioned in Section 6, SpeechView’s LipC software sits on a computer connected to a telephone that is receiving sound. The software listens to the sounds arriving at the telephone, identifies phonemes, computes visemes for the phonemes, and both outputs the sound on the computer’s speaker and displys the lips on the computer’s screen. The telephones sound output is disabled, and the software delays the sound so that it can be played synchronized with the displayed visemes that it calculates. Ironically, when the bandwidth of the signal to telephones and cellular telephones will be high enough for alive-action video to be transmitted synchronized with the sound, the attractiveness of the SpeechView approach might decrease. However, as mentioned in Section 6.3, because of the additional signs LipC shows on the throat, nose, and cheeks to help distinguish phonemes that share a viseme, the LipC generated lipreadable lips carry more information than do natural human lipreadable lips; thus, the HI individual with LipC may indeed prefer LipC's artificial lipreadable lips to equally available alive-action video.

Voice recognition is improving steadily to the point that there are products that can be taught to translate one user's voice into ASCII text. Perhaps in the near future, software will be able to translate an arbitrary voice or a voice in a set of hundreds of previously training voices into ASCII text. When such technology is available, it should be utilized to provide real-time captioning of voices, both on TV and in voice-based user interfaces. Even if the accuracy were not perfect, but were only 95%, it might be usable by the HI. We are quite used to sloppy, slightly delayed captions produced by human courtroom-style stenographers in real time during alive news and sporting event broadcasts. The mistakes are plentiful and sometimes amusing. Most often the mistake is to a sound-alike sequence of words, e.g., "eye deal" instead of "ideal", and the listener has to listen to herself speak the words mentally. My feeling is that the technology will be no worse than the current real-time captioning.[13]

Notes

[1] Originally, all telephones were rented to the customers by the owning telephone companies, who had the obligation to keep all the equipment under repair, even the telephones in customers' homes. Thus, the telephone companies had a strong incentive to build telephones to be indestructible, and it seems that the telephones *were* indestructible. Nowadays, customers *buy* their telephones. It seems that telephone companies now have a strong incentive to make the telephones more fragile so that customers have to buy new ones occasionally.

One telephone technician explained to me that in the old days, the volume range supported by the Western Electric amplified telephone was much larger than that given in the official specifications and was thus much larger than the range offered by the amplifier's dial. Thus, when the dial is at its maximum, the telephone is delivering only a fraction of the volume that is possible and there is very little distortion. Nowadays, says the technician, the volume range supported by the average amplified telephone is much closer to the specified range. Thus, when the dial is at its maximum, the telephone is delivering at nearly its capacity, and there is lots of distortion.

[2] I understand that voice-directed menu selection is universally disdained, even by non-HI users of telephones.

[3] To avoid heavy usage of "he or she" as a third person singular personal pronoun, this paper alternates, on a section-by-section basis, the gender of the arbitrary persons introduced by quantifier equivalents.

[4] These claims come from personal experience, observations of other HI individuals I have met, and observations of my fellow patients in therapy situations over the years.

[5] A TTY unit is a keyboard plus modem that communicates directly with other TTY devices over telephone lines using the 5-bit Baudot code at 150–300 baud (Williams, 1998). Consequently, it is incompatible with ASCII and the e-mail world. Thus, TTY users form essentially a closed world. However, with proper software, a computer with a modem could connect with a TTY unit (TAP Program, 2002).

[6] As one who reads lips, I can say that the snapshots are well chosen, as I can instantly recognize the possible

phonemes for each of the visemes!

7. A *viseme* is the pattern of lip, teeth, tongue, and other facial movements for one phoneme.[8] In practice, the lip movements are so dominant in and so characterizing of a viseme, that reading visemes is called lipreading. I, for one, ignore everything but the lips.

8. A *phoneme* is a unique sound for one letter or diphthong. One letter, e.g., “c”, may have more than one phoneme, e.g., as in “car”, as in “cedar”, and as in “Mercedes”. Conversely, several letters may share the same phoneme, e.g., the “s” in “see” and the “c” in “cedar”.

9. The proper technical term is “plosive”, but “explosive” is more descriptive for the benefit of the lay person; Within the ch–sh–j viseme for English, “ch” is plosive and “sh” is non-plosive, and definitely, “ch” is more explosive than “sh”!

10. For reasons beyond the scope of this paper, I believe that teaching signing or even signing and speaking is the worst thing that can be done to a HI person. He learns to sign, does not learn to speak, and can interact only with other signing people. Not teaching signing leaves the HI person no choice but to learn to read lips and to utilize the residual hearing he has. He does so with no more effort than hearing people learn to understand spoken language and than HI people learn to sign.

11. This way, I avoid having to deal with those #%&! telephone solicitors entirely. Although, I wonder how motivated these solicitors are to connect with people; despite the instructions on the recording, not one of those solicitors has been willing to take the effort to send me a fax!

12. As a consequence, the HI community is cut off even more from the rest of the world, which has gone ASCII and into bandwidths in the thousands of baud.

13. Of course, for previously recorded TV shows, series, etc., it is possible to do perfect and synchronized captioning. Since the code used by the closed-captioning system is ASCII, often an ASCII rendition of the script is used. In this case, sometimes the captions do not agree with what is actually said. The actor said something that meant the same thing and the director accepted the change. However, the captions remain a copy of the script.

References

Berry, D.M. 2001. Requirements for Maintaining Web Access for Hearing-Impaired Individuals. *Third International Workshop on Web Site Evaluation (WSE’01).* Florence, Italy, pp. 33–41, Also at http://se.uwaterloo.ca/~dberry/FTP_SITE/reprints.journals.conferences/WSE_paper.pdf.

Boyd, M. 2002. Signs that Tessa Can Help the Deaf. From *Sunday Star* 10 November 2002, Read 12 November 2003, Keratan Akhbar & Majalah, e-pek@k, http://www.epekak.net.my/akhbar/2002/11_tessa.htm.

Chisholm, W., Vanderheiden, G., and Jacobs, I. 2001. Web Content Accessibility Guidelines 1.0. *ACM Interactions*, VII(4):35–53, Also at http://www.w3.org/TR/1999/WAI-WEBCONTENT-19990505/.

Comet, M.B. 2003. Lip Sync - Making Characters Speak. Read 12 November 2003, http://www.comet-cartoons.com/toons/3ddocs/lipsync/lipsync.html.

EASI 2003. Equal Access to Software and Information. Read 12 November 2003, Rochester Institute of Technology, http://www.rit.edu/~easi/.

Fainchtein, I. 2002. Requirements Specification for a Large-Scale Telephony-Based Natural Language Speech Recognition System. Master’s Thesis, School of Computer Science, University of Waterloo, Waterloo, ON, Canada.

IBM 2003. ViaVoice. Read 12 November 2003, http://www-3.ibm.com/software/voice/viavoice/.

ICDRI 2003. International Center for Disability Resources on the Internet. Read 12 November 2003, http://www.icdri.org/.

Leavitt, N. 2003. Two Technologies Vie for Recognition in Speech Market. *IEEE Computer*, 36(6):13–16.

Lincoln, M. 2001. TESSA. Read 12 November 2003, University of East Anglia, School of Computing Sciences, Norwich, UK, http://www.sys.uea.ac.uk/˜ml/tessa/Tessa.html.

Marcus, A. 2003. Universal, Ubiquitous, User-Interface Design for the Disabled and Elderly. *ACM Interactions*, X(2):23–27.

Martin, G.C. 2003. Extended Preston Blair Phoneme Series. Read 12 November 2003, http://www.garycmartin.com/phoneme_examples.html.

Ross, M. 2002. Quality in Web Design for Visually Impaired Users. *Software Quality Journal*, 10(4):285–298.

Signtel Inc. 2003. Deaf In America. Read 12 November 2003, http://www.signtelinc.com/dia-1.htm.

Signtel Inc. 2003. Opening the Doorways of Communication. Read 12 November 2003, http://www.signtelinc.com/.

SpeechView 2003. Your Link to the Hearing World. Read 12 November 2003, http://www.speechview.com.

SYS Consulting Ltd 2003. The Tessa Project. Read 12 November 2003, http://www.sys-consulting.co.uk/web/projects/project_view.jsp?code=TESSA.

TAP Program 2002. Can I Use a Regular Computer Modem to Call a TTY?. Read 12 November 2003, Gallaudet College, http://tap.gallaudet.edu/faq1.htm.

Third Wish Software 2003. Magpie. Read 12 November 2003, http://www.thirdwishsoftware.com/magpie.html.

ViSiCAST 2003. TESSA. Read 12 November 2003, http://www.visicast.co.uk/news/Tessa.html.

Wang, J. 2003. Human-Computer Interaction Research and Practice in China. *ACM Interactions*, X(2):88–96.

Watchfire Corporation 2003. Welcome to Bobby. Read 12 November 2003, http://bobby.watchfire.com/bobby/html/en/index.jsp.

Williams, N. and Harkins, J. 1998. TTY Basics. Read 12 November 2003, TAP Program, Gallaudet College, http://tap.gallaudet.edu/TTY-Basics.htm.