In "Proceedings of the International Conference on Cybernetics and Society", IEEE Systems, Man and Cybernetics Society, Atlanta, GA., October, 1981, pgs.393-397.

THE INTELLIGENT EAR

A GRAPHICAL INTERFACE TO DIGITAL AUDIO

Christopher Schmandt

Architecture Machine Group

Massachusetts Institute of Technology

ABSTRACT

The "Intelligent Ear" is a digital dictaphone driven by an interactive graphical interface. This system experiments in the use of cross media mapping, between audio and its visual representation, to facilitate man-machine interaction in sound editing tasks.

The Ear's intelligence includes limited keyword recognition and display of amplitude, i.e. phrashing data. A color video display is capable of communicating this content simply but meaningfully, by mapping temporal audio events into spatial visual cues. The experience of another area of information processing, screen oriented text editors, contributes to an easily used interface through which the user can always visualize the current state of the edited sound.

The "Intelligent Ear" is an interactive graphical interface to a digital audio recording and playback system. The "Ear" is both a display oriented, content sensitive listening device and a "screen editor" for recorded sounds. It consists of a digital audio recording and playback system coupled to a speech recognizer; display and control is via a color raster scan television monitor overlayed with a touch sensitive surface. This system attempts to facilitate audio related tasks by providing a human interface to digitized sound using computer graphic techniques.

The Graphical Audio Interface

The design emphasis of the Intelligent Ear is on the interface, via graphics, to audio communications. We are attempting to show that a smart, and particularly a highly interactive, display has the potential to revolutionize control of otherwise non-graphical media. A key point is the Ear's intelligence; not only does it allow access to sound data, it also attempts to understand it.

As a listening device, the Ear digitizes a conversation or dictation and stores it on magnetic disk. It later scans the recording for the occurance of selected keywords. The recorded audio is then displayed graphically via a standard raster-scan color frame buffer. Sound amplitude modulates both height and color of a waveform representation of the recording drawn on the display. The keywords which have been recognized in the speech are written on the monitor below the appropriate location in the waveform representation of the sound.

Interacting With Digital Audio

The Intelligent Ear can be used in a variety of ways depending on the particular application. In the most general case, we assume that a conversation or dictation has been pre-recorded at some time and is now to be either reviewed or edited, perhaps by the same speaker, perhaps by another. At one extreme the Ear can be used as a "minute taker" at a meeting; some time later, it displays graphically the conversations of the meeting, with different speakers shown in different colors, and keywords noted in the meeting highlighted textually. At the other extreme, the Ear is a dictation device which allows easy and clean editing of memos, letters, etc; again, keywords can be displayed for more intuitive understanding of the speech waveform displayed by the editor.

The Ear's interface consists of a representation of the recorded speech, and touchable "buttons" similar to those of a conventional tape recorder, but with powerful additional editing functions. The sound is displayed much as on a waveform monitor or graph; lines across the screen vary in height as well as intensity as a function of the amplitude of the audio signal. Horizontal (i.e. time) resolution is adjusted so that pauses between sentences and in some cases between words are clearly visible, allowing easy access to clean edit points.

The work reported herein was supported by the Cybernetics Technology Division of the Defense Advanced Research Projects Agency, under Contract No. MDA-903-77-c-0037

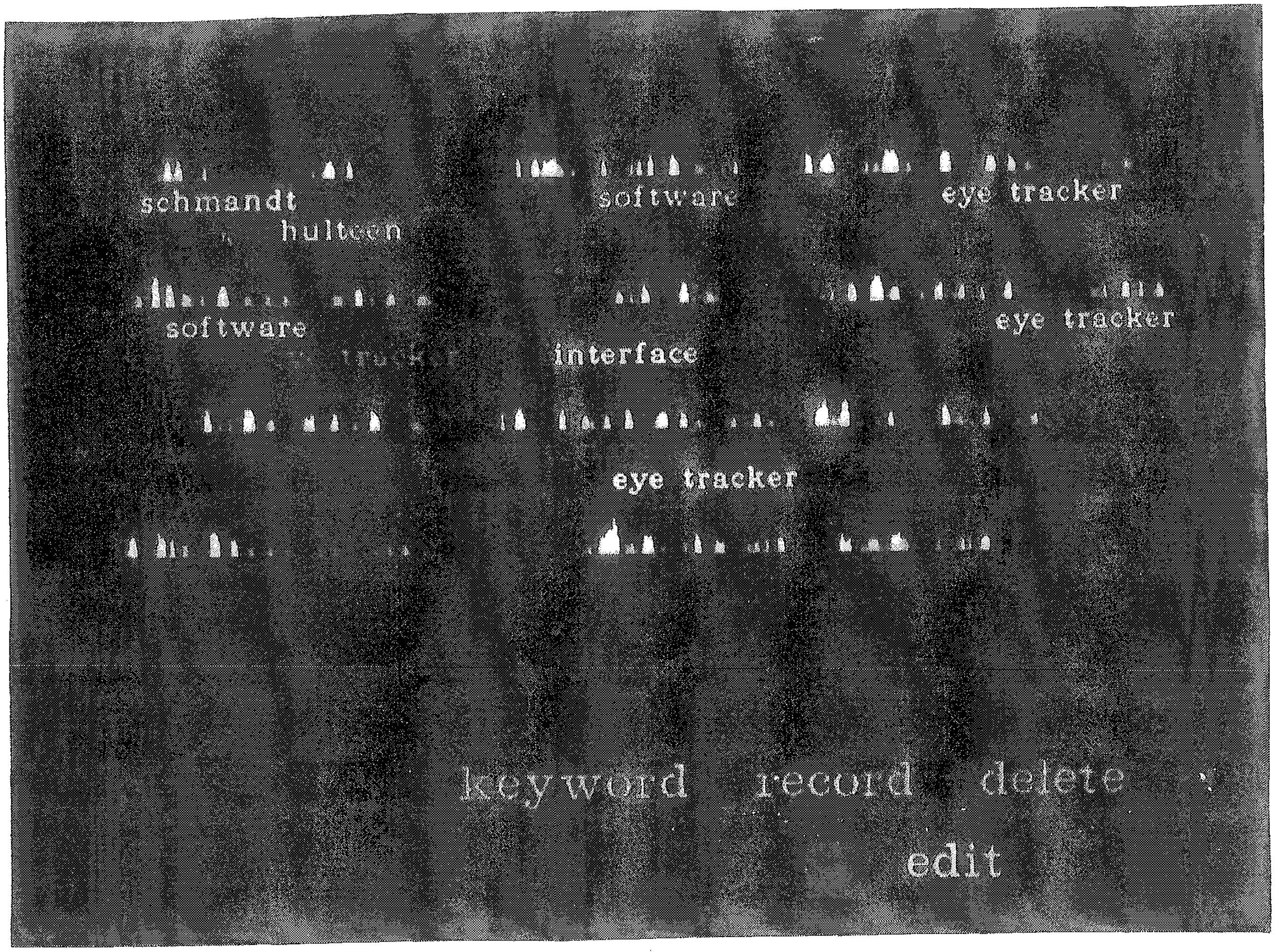

Figure 1. The Intelligent Ear. This photograph was taken from a standard television monitor; the graphics use color in several areas. The screen is touch sensitive, and the words across the bottom are "buttons" which control editing.

A "sound cursor", a bright colored rectangle, is positioned by touching the desired point on the sound waveform; it indicates where to make an edit or the point from which to start playing a sound (see figure 1).

By touching a "play" button at the bottom of the screen, the sound starts playing from the current position of the sound cursor. While the sound is playing, the associated waveform changes color in sync with the audio, i.e. the user can visually identify which part of the recording is playing, a useful editing aid. The sound plays until a "stop" button is touched, or the end of the sound is reached.

Note that there is a direct relation between the (x, y) point touched on the screen and a time offset into the recording. The waveform changing color in space while, and quite in sync with, the audio playing in time creates a powerful spatial association between the two media. This cross-media link is a prerequisite for an interface which allows immediate and intuitive interaction (1, 2).

Selected keywords which have been detected are written under the appropriate point on the sound waveform display. Again, this allows a higher degree of visual perception of the content of the recording. Although user interaction with the graphical interface brings nearly immediate response, keyword recognition is not a real time process, taking up to twelve times the length of a conversation to analyze it. The graphical interface communicates the degree of certainty for recognition of each word in the intensity with which the word is written; the Ear displays words it is more confident of brightly, and ones with less confidence in less visible greys.

As an editor, the Ear allows both insertion and deletion of audio material. A "record" button allows the user to speak into the sound document at the current position of the sound buffer. Similarly, a "delete" button proceeds to erase whatever sound lies between the next two points on the sound waveform depiction the user touches. The editor analyzes the audio signal around the various edit points to make a smooth edit; it's intelligence includes finding sentence or phrase boundaries.

Whenever an edit is made, either a recorded insertion or deletion, the graphical display is updated to show the latest change. Although keyword recognition cannot be done on insertions in real time, the waveform representation changes quickly, and the new version of the sound remains editable. Since editing is buffered (see below), extra red and green buttons are provided to save or restore the edit buffers.

A Display Oriented Editor

As an editor, the Intelligent Ear makes use of many concepts which have gained popularity, with good reason, among users of computer text editing systems. With the advent of inexpensive computer terminals with display capabilities, a number of highly interactive editors have been written, and often are one of the most widely used programs on a computer system (3, for example). Two specific features of such editors have been deliberately incorporated, display orientation and buffered edit operations.

A display oriented text editor is one in which a text file is continuously displayed in its current state. Typing control functions which inserts, delete, or alter text update the terminal screen immediately. This instant feedback is invaluable for keeping track of the current state of the document being edited, and of the editor's recognition of the typist's intent. The same idea is incorporated into the Ear for identical reasons. Whenever the sound is edited, the display of the sound amplitude, keywords, etc. is quickly updated. Cursor positioning, again as with a text editor, shows where in the sound document the edit is to occur. This makes learning the operation of the Ear a quick, intuitive process, and insures that the user is more likely to make the edits he/she actually wants.

In addition, all edits are buffered. Two copies of the sound are maintained, and a single edit is effected to only one of them. The sound can be thought of as a list of buffer offsets to play sequentially. In the case of a deletion, the sound is played up the beginning of the deletion, then skips to the audio section following the edit. In a recorded insert, the sound is played up to the insertion, then a separate record buffer is played, then back to the original sound. Touchable buttons then copy the edited sound into the second edit buffer, so it will play as edited in a single linear playing. This feature gives a chance to review every edit by playing it before any audio date is permanently changed.

Keyword Recognition

The main barrier to speech recognition from normal conversational speech (as opposed to the clearer enunciation and pauses between words usually associated with speech input devices) is the blurring together of a number of words into coarticulated phrases (4). Although the Nippon Electron Company (NEC) DP-100 is remarkably successful analyzing connected words, it can process a maximum of 2.4 seconds of continuous input, without pauses. These pauses tend to be absent in ordinary conversational English.

Using the flexibilty of a computer contolled digital audio system we devised a scheme to overcome the continuous input limitations of recognition hardware by playing back small segments of the recorded sound with waits between each segment for the DP-100 to perform its analysis. The recording is divided into sections which are played sequentially. Since words may be chopped at the segment boundaries, successive audio windows must overlap. What one actually hears during the keyword analysis of the recording is a moving window of sound, with pauses between each play, and overlap between each segment.

Performance falls within a wide range depending on whether the recording is of a single person dictating or a multi-speaker conversation. Performance tends to be a trade off; as we lower the level of confidence required for acceptance, the number of correct recognitions increases, but so do the false guesses. Allowing one false guess per minute (reasonable in examples where keywords occur quite a bit more often) we have been able to recognize 25% to 75% of the keywords in our experiments.

Several schemes have been developed to give higher recognition. A multi-pass analysis is used; the same recording is played to the speech recognizer three times, with varying window lengths and overlaps. A majority poll is taken between the passes to determine what counts as a detection. This reduces false guesses as the signal (correct recognitions) tends to remain more coherrent across passes then the noise (false guesses on different passes tend to chose different words). This multi-pass algorithm roughly doubles keyword detection performance, and has yielded results of between 40 and 100% of correct recognitions at the one false guess per minute rate.

It is important to consider the impact of imperfect keyword recognition on the task at hand. In fact, several graphical techniques increase the utility of our detection algorithm. The first takes advantage of the fact that our keyword detection algorithm associates a confidence value with each recognition. The text of keywords is written below the sound waveform display in a manner which both indicates the confidence of recognition and simultaneously allows user filtering of these results. The more confident we are of a keyword's existence, the brighter it is displayed in a greyscale of possible text intensities. Since the less confident guesses are in fact more likely to be incorrect, they are written with faint text so they can be easily ignored. A user can quickly associate a trait such as brightness with the Ear's certainty that the word really exists.

The second approach allows easy user feedback into the keyword discrimination process. Two touch sensitive "buttons" are labelled "verify" and "erase". Touching either button, then the text for a keyword, causes that keyword to either turn full white (verify) or vanish (delete). From this point on, the Ear notes either 100% certainty of the existence of a keyword, or ceases to display it; as this modifies a database associated with the sound recording, future uses will reflect this modified confidence value. Thus the effective intelligence of the Ear can be boosted by allowing the user to contribute his/her own keyword detection.

<u>Hardware Configuration</u>

User interaction with the Ear is via a touch sensitive color display. Graphic images are drawn on a Ramtek 9300 frame buffer; this provides 9 bits per pixel of conventional raster scan video. The monitor screen is covered with a clear

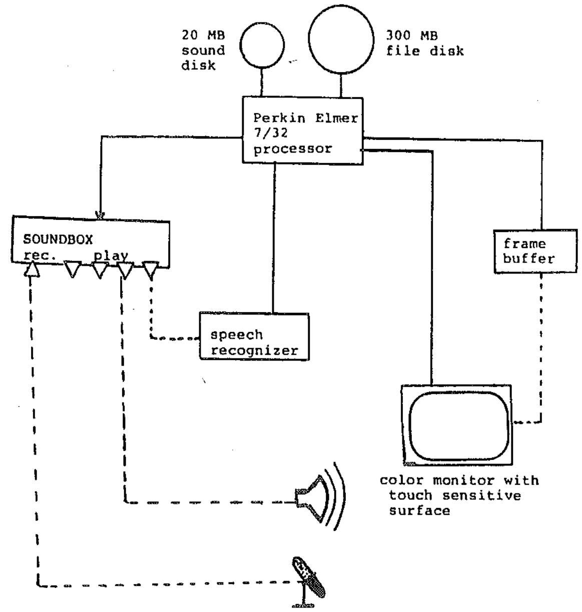

Figure 2. The hardware configuration. Solid lines indicate digital data paths; broken lines are analog audio or video.

touch digitizing plastic surface manufactured by Elographics. All software and device interfaces inhabit a Perkin-Elmer 7/32 minicomputer with 512 Kbytes of memory. (see figure 2)

At the heart of the Intelligent Ear is the Laboratory's own design digital audio recording and playback system, called the Soundbox (5,6). The Soundbox records and plays audio with a useable bandwidth of approximately 3.8 KHz, at eight bits of resolution; this produces audio of approximately telephone quality. Such recording fidelity is quite sufficient for experimental work involving voice bandwidth audio even though the human voice contains frequencies up to about 8 KHz.

Software drivers for the Soundbox move blocks of digitized audio data between a dedicated 20 megabyte magnetic disk and record and playback buffers. Sounds are stored on a disk in a file system which allows conventional data management operations: deletion, concatenation, copying, etc. Up to four sounds ("voices") may be played simultaneously. Voices may also be preloaded with audio data from disk so they may be played sequentially with no audible pause between them, a feature we use for smoothly playing sounds located in several editing buffers.

A speech recognition system is connected to the audio output of the soundbox as well as interfaced digitally to the Ear's computer. Recognition is accomplished by a Nippon Electric Company DP-100 Connected Speech Recognition System (7). This unit is capable of recognizing up to 120 selectable words from continuous human speech. Connected speech is an important requirement in this application since we are analyzing normally spoken conversations without pauses between words. The DP-100 will process up to five words without pauses; for our keyword recognition operation the Soundbox breaks the speech into arbitrary phrases rather than trying to search for word boundaries, a formidable task. Another important software feature of the Laboratory's version of the DP-100 is that in addition to returning to the host computer the word it has recognized, it also returns a "confidence value" indicating how close a match was found with that detection.

Summary

The Intelligent Ear concentrates on graphic techniques at the merger of several disciplines. An interactive color display is a highly useable interface to perusing and editing digital audio data. The Ear's intelligence includes limited keyword recognition and display of amplitude, i.e. phrasing data, about the sound. A color video display is capable of communicating this intelligence simply but meaningfully. Finally, the experience of another area of information display, screen oriented text editors, contributes to an easily used and very practical editing system.

Acknowledgements

Both Eric Hulteen and Mark Vershel, graduate students at the Architecture Machine, have aided this project immensely, both in sharing the programming tasks and in contributing creative ideas into the Ear's design and keyword recognition algorithms. Their assistance is more than appreciated.

References

1. Negroponte, N. Augmentation of human resources in command and control through multiple media man-machine interaction. MIT Architecture Machine Group, ARPA Report, 1976.

2. Bolt, R.A. Spatial-data-management, MIT Architecture Machine Group, DARPA Report, 1979.

3. Ciccarelli, E. An Introduction to the EMACS Editor, MIT Artificial Intelligence Laboratory Memo #447, January 1978.

4. Reddy, D.R. Speech recognition by machine: a review. Proceeding of the IEEE, 64, 4 (April 1976), 501-531.

5. Hurd, Jonathan A. An Interactive Digital Sound System for Multimedia Databases. S.B. Thesis, MIT, June 1979.

6. Vershel, Mark. The Contribution of 3-D sound to the Human-Computer Interface. M.S. Thesis, MIT, June 1981.

7. Kato, Yasuo. Words into action III: a commercial system. IEEE Spectrum (June 1980), 29.