THE DESIGN OF A SPEECH ANALYSIS WORKSTATION

JOHN M. CRUMP

Kay Elemetrics Corp
12 Maple Ave.
Pine Brook, NJ 07058 USA

ABSTRACT

The development of a speech analysis workstation is presented. The problems and challenges in acoustically analyzing speech signals are discussed. A system was developed to provide the digital acquisition and analysis of speech with all of the features typically required in acoustic phonetic research.

INTRODUCTION

Speech has been acoustically analyzed by a wide assortment of instruments including oscilloscopes, spectrographs, and numerous computer based systems. Typically a computer system requires a number of peripherals to analyze speech. These peripherals may include input modules with A/D and anti-aliasing filter, graphic boards and special printers. High speed array processors or special digital signal processing boards may be added to boost processing speed. Software to analyze the stored signal is typically purchased commercially or developed by researchers.

The recent availability of general purpose digital signal processing chips, inexpensive digital memories and personal computers has provided the technical capabilities for the development of a powerful workstation designed for the analysis of speech. A system can now be developed with the advantages of a spectrograph (e.g. Sona-Graph and SSD), an oscillograph (e.g. Visicorder), a feature extractor (e.g. Visi-Pitch), and a general purpose computer (e.g. VAX with DSP software).

DEFINING A SPEECH WORKSTATION

Before the development of a speech analysis workstation is started, it is important that the analysis requirements of the users are clearly understood. Speech is analyzed by many different professionals for many different reasons. A phonetician may have different needs than an speech language pathologist. Any workstation designed for speech analysis must take these different requirements under consideration. The common elements for most speech analysis are reviewed as follows:

Input

The aliasing portion of a signal must be filtered before the signal is digitally stored. Low-pass filtering is the process of eliminating the high frequency components which will create spurious spectra in the analysis. Providing adequate anti-aliasing filters is a difficult, and often overlooked, problem especially if the user changes sampling rates to perform different analysis tasks. For example, the analysis of vocal behavior (e.g. perturbation measurements) requires very high sampling rates to achieve high timing accuracy. Sampling rates as high as 50-100kHz may be required. Anti-aliasing filters at these sampling rates are quite different from filters at slower sampling rates.

Sampling frequency must be variable and should exceed the 50 kHz sampling rate required in some applications.

Psycholinguistic experiments and phonetic transcription require a system which can store and playback speech at high fidelity. High fidelity playback requires high sampling rates. If the workstation is to be used to acquire and define a phonetic library the speech signal requires a deep dynamic range and excellent frequency response. Dynamic range should be above 70dB and sampling rates above 50kHz. The speech signal storage should be sufficient to store at least one paragraph of speech sampled at high rates.

All of the above requirements are very important because there is a general requirement for instrumentation to simply acquire, filter, amplify/attenuate, A/D, D/A and buffer speech signals for input to computers for further analysis. A speech workstation should be able to excel in this limited but important function.

From the requirements explained above the following criteria for input and signal storage were developed:

Sampling rates: Variable with samples up to 80kHz

Dynamic Range: 12 bits or >72 dB

Low pass filters: Automatic with sampling selection, 120 dB/octave, preferrably digital filters.

Signal storage: At least 40 seconds sampled at 20kHz. This requires 2 Mbytes of memory.

Displays

Graphically, speech has tradionally been displayed as a waveform, a spectrogram, a power spectrum (frequency vs. power) or as tracings of speech parameters. A speech workstation should be able to present these four standard displays clearly and crisply. Speech analysis also typically requires timing and frequency measurements. Various feature extraction techniques such as LPC analysis has also proven itself a useful tool. Integrating these various approaches in the analysis of speech would be especially useful. For example it would be useful to superimpose color LPC extracted formant values on a wide band grey scale spectrogram. Depending on the analysis task it would also be desireable to be able to rapidly switch analysis formats to find the type of display most revealing of the characteristic under investigation.

A workstation should allow a wide range of display options which can be quickly performed (less than 2 seconds). This will help users quickly re-analyze the stored data to find the most revealing display of the aspect of interest. Time resolution of waveform displays must facilitate the measurement phenomenom of both very short and long duration. Timing accuracy should be as fine as each data point of memory for resolution of 0.01 milliseconds. Spectrograms must include a selection of analysis filters for the fine time and frequency resolution frequired for the effective formant display of low and high pitch voices.

Real Time Performance

Real time analysis is valuable for a number of reasons, some obvious and others not so obvious. The faster the analysis is performed the less waiting for the user. If the user can quickly re-analyze data he or she is more likely to explore various analysis modes to find the most revealing method. In any clinical setting real time analysis is usually a requirement.

The other advantage of real time analysis is that the data can be monitored during input and analysis. Systems, which batch analyze data, require the user to first store data and then analyze. Speech is such a dynamic signal that unless the input can be monitored during input it is very difficult to acquire the signal without overloading during transient peaks or underutilizing the full dynamic range. One solution is to use input systems with very deep dynamic range (>90dB) which require 16 bit A/D and extremely good low noise input circuits and anti-aliasing filters. These systems are very expensive.

For many applications it is important to monitor the analysis in order to select the correct data for analysis. For example if the researcher is investigating an acoustic phenomenon which is clearly displayed spectrographically, but is difficult to hear, real time capability allows the user to scan the input speech signal to select the appropriate segment.

Some systems will analyze in real time, but can not simultaneously store the speech signal. This is obviously undesireable because the user must re-enter the signal to re-analyze. A true real time system must be able to simultaneously low pass filter, acquire, store to memory, analyze and display in real time.

Graphic Resolution

As mentioned above the graphic displays are an important component in any speech analysis workstation. High resolution graphic displays are technically difficult. Typical microcomputers video graphic standards fare not good enough to replicate the display resolution of even 1950 style hard copy spectrograph. The selection of grey scales available are insufficient to display spectrograms. The fine timing and frequency measurements require a more robust display standard with more than 32 shades of grey for each element and a display resolution of at least 640(H) x 480 (V). Hard copy resolution must match the standard set by the commonly available hard copy spectrographs. A color display would also be useful to display speech parameters (such as LPC extracted formant frequencies) and grey scale spectrograms simultaneously. Color is also required when multiple traces are displayed.

Interface to Computers

A speech workstation should be able to operate inside a microcomputer, or be easily interfaced to microcomputers. For a number of reasons discussed in more detail in another section of this article currently available microcomputers can not become practical speech workstations. Despite these limitations inexpensive microcomputers can serve valuable functions if interfaced to a speech workstation. The availability of inexpensive file management, data storage and software complement the analysis and display power of a speech workstation. An interface to these microcomputers should be very fast to facilitate rapid exchange of data files and to increase the utility of the speech workstation as a data acquisition peripheral.

Programmability

The rapid advances in digital signal processing of speech necessitate that a speech workstation can be updated to apply new algorithms to speech analysis. Often users are only interested in a single speech analysis measurement and may require adjustments to currently available programs to best extract this information. It would be desirable for the user to be able to change programs and a requirement that the vendor can upgrade without using software rather

than hardware replacement.

User Friendly

A speech analysis system will often be used by speech scientists, speech language pathologist and phoneticians who may not be instrument oriented or computer specialists. They also may only perform acoustic analysis infrequently in their work. In this working enviroment, it is important that a speech workstation is easy to use. The system should be menu driven and methods of analysis/display should be electronically storable and retrievable so that users can repeat analysis methodology exactly.

In a teaching enviroment acoustic analysis tools are often used to teach students about acoustics. It would be useful for a workstation to be designed to facilitate this task by storing precisely repeatable acoustic analysis experiments.

Dual channel Capablility

Speech is often investigated in conjunction with other physiologic signals. A speech workstation should be able to operate in dual channel mode to analyze electroglottograph, airflow, accelerometer and other signals of interest in conjunction with the speech signal.

Affordability

Price and performance have obvious tradeoffs in any development but a speech workstation can not be beyond the reach of most speech scientist no matter how wonderful the product is.

EXPLORING THE AVAILABLE TECHNOLOGY

Once the outline of the features and specifications were established the commercially available technology was investigated to determine the best approach to accomplish the design criterion. One approach which was considered in detail was the packaging of the hardware/software for this workstation inside a standard microcomputer. In this configuration the hardware would plug into the backplane of the DEC Q- Bus, the IBM-PC bus or directly connect to a high speed port of other computers. DEC , IBM-PC ATs, Amigas, Apollo, Sun, Masscomp, MacIntosh and others were evaluated.

Incorporating the workstation in these comon computers was rejected for technical and/or cost considerations. The widely available inexpensive computers (IBM-PC, Amiga, MacIntosh etc.) were not powerful enough even with added hardware. The technical limitations of inexpensive microcomputers to perform as a speech workstation are as follows:

1. The bus of microcomputers has a very limited bandwidth and it can not, therefore, acquire signals at the sampling rates required for many speech analysis tasks.

2. The bus and DMA capabilities of microcomputers do not allow the simultaneous transfer of data from input board to memory, input board to analysis module, analysis module to display memory. It can not, even with the addition of graphics, input and digital signal processing boards do true real time acquisition, analysis and display.

3. Most computers have insufficient memory available for signal storage. As noted previously at least 2Mbytes of signal stoage are required in addition to 512K bytes of digital signal analysis work space and 384kK bytes of display memory.

4. The digital signal processing speed is at least 100 to 200 times too slow for real time analysis. Accelerator boards can be added but the speed is still insufficient for a robust system.

5. The highest standard graphic standards on microcomputers are not able to display spectrograms with enough resolution in time (horizontal), or sufficient grey scale. Many computers restrict the user to specifc color selections because the video controller can only turn on or off each RBG output guns. This restriction does not allow the subtle variation of hue or grey scale necessary in some applications.

The more powerful systems are costly and not widely or consistently available for many potential users. Even these more powerful systems (VAX etc) are too slow for the real time digital signal processing required. Array processors would need to be added to achieve real time performance and, in some cases, the system architecture can not transfer data blocks at the required rates.

These technical and cost considerations aside, it must still be emphasized that it is important to have high speed interfacing between the standalone speech workstation and the widely available IBM-PC type microcomputer and VAX minicomputers. High speed interfacings eliminates the need for the workstation to include its own disk drives and allows access to available DSP software and previously digitized data.

DEVELOPING THE WORKSTATION

The result of the exploration has led to a standalone system based on a common microprocessor, powerful digital signal processing integrated circuits, high resolution graphic displays and high speed DMA capabilities. The digital signal processing chip selected was the 32020 from TI (Texas Instruments). Two 32020s are used to further increase the processing speed to ten million instructions per second. The 32020 are capable of many parallel operations and include a fast single-instruction multiply operation. These features are extremely useful because of the repetitive nature of the instructions and the many multiplications required in digital signal processing. These features combine to provide digital signal processing speeds equivalent to over 50 million instruction per second in a general purpose computer. These chips were also selected because of the upward migration path TI has produced with the 320C20 and 32030.

Two separate buses for data acquisition and analysis were used. This "extra bus" and special high speed DMA chips were used to facilitate high speed data transfer between the different system modules (A/D to memory, memory to DSP circuits, DSP circuits to graphic circuits and DSP circuits to printer). These DMA chips allow a 4 Mbyte/sec transfer rate. The system management is performed by a Motorola 68000 and the system architecture has been defined to include up to 8 Mbytes of RAM and 2 Mbytes of PROM.

The graphic resolution required for both the real time display monitor and hardcopy were the most difficult to achieve. The system was designed with a graphics controller, high speed video DRAMs and a special monitor to provide graphic resolutions of 640 x 480 with 256 values of color and/or grey scale for each pixel. The sytem allows simultaneous grey scale and color displays because the monitor and video driver are capable of both analog and digital display. Because the extensive graphics routines required in a speech workstation can not processed quickly through the CPU the graphics hardware was designed to perform most of the graphics displays without CPU intervention.

The hardcopy print capability is based on a new thermal printer and this print quality matches the quality of Kay sonagrams™ which have become the standard for spectrographic display. The printer produces true (not imitated with a collection of dots turned on or off) grey scale at 120 dpi.

The system requires multiple processing modules to achieve the speed and performance required. The relatively slow CPU is relieved of virtually all of the processing, except for controlling the other modules.

The system meets all of the criteria set above for a speech workstation. It can not be programmed by a novice and is, therefore, limited to the programs available from Kay or programs developed by programmers familiar with the TI 320 code. There are over 320 design teams working with this chip according to TI. How many are working in the speech field is not known but the TI320 family represents over 65% of the digital signal processing chips sold in 1986. It has become a standard for digital signal processing development and there are numerous plug-in boards for computers designed for 320 code development. Kay has developed a series of programs to implement all of the features discussed in the the section "DEFINING A SPEECH WORKSTATION". Along with the development of numerous speech analysis programs continuing at Kay other groups, including the University of Victoria's CSTR (Centre for Speech Technology Research), are working on LPC analysis / modification/ synthesis programs. Kay will commercialize the programs developed by CSTR. The system has all of the programs stored on a large PROM board to facilitate updates as the science of digital signal processing develops.

When interfaced to computers the speech workstation can also be used for input, speech selection and buffering, display and grey scale printing. Users can then use the programming tools available on their computer for other digital signal processing or file management programs.

SUMMARY

The system succeeds in meeting the design criterion for a general purpose standalone workstation. State of the art technology and multiple processing modules were required to meet this criteria. To facilitate its utility as a peripheral to common computers such as the IBM-PC and DEC VAX, software is being written to exchange data and allow these computers to easily use the powerful graphics, data acquistion and digital signal processing capabilities of the standalone speech workstation.