SPEECH COMMUNICATION GROUP

WORKING PAPERS VOLUME IV

May, 1984

Research Laboratory of Electronics
Massachusetts Institute of Technology

Correspondence regarding this publication should be addressed to:
Charla Scivally, Speech Communication Group
Massachusetts Institute of Technology
77 Massachusetts Avenue
Building 36, Room 541
Cambridge, Massachusetts 02139

7. THE NEW M.I.T. SPEECHVAX COMPUTER FACILITY

D. H. Klatt

7.1. Introduction

The Speech Communication Group has recently acquired a new VAX-750 computer facility to take over speech analysis and synthesis functions formerly performed by the PDP-9. This brief note is addressed to others facing a similar need to upgrade a computer facility for speech research. It identifies some of the issues we faced and our chosen solutions. Another purpose of this note is to describe software under development.

The PDP-9 computer was a well-designed single-user speech research facility. It possessed a home-made fast display for visually instantaneous waveform scrolling, and a hardware filter bank that permitted rapid updating of a crude spectral representation while scrolling the waveform (Henke, 1968; 1969). These functions are not easy to duplicate with presently available standard computer hardware — at least when budget constraints apply — especially when additional constraints such as an ability to simultaneously handle several users are desired.

The hardware that was selected to satisfy the computing needs of the former users of the PDP-9 is shown in Figure 7-1. It was all purchased from Digital Equipment Corporation, in part so that a single service contract could be used to avoid conflicts when it is not certain which piece of hardware is causing an obscure software problem, and in part because of the large discount provided by Digital's external research grant program.[1] Software is being written in C under the VMS operating system that is provided by Digital. We considered using Berkeley Unix as an operating system, but at the time that a decision had to be made, there was no satisfactory Unix driver available for the LPA a/d and d/a converters. Otherwise, the relative advantages and disadvantages of the two operating systems would have made for a very difficult, and hence rather arbitrary, decision.

---

[1]We are also grateful for the additional funds provided by the office of the M.I.T. Provost.

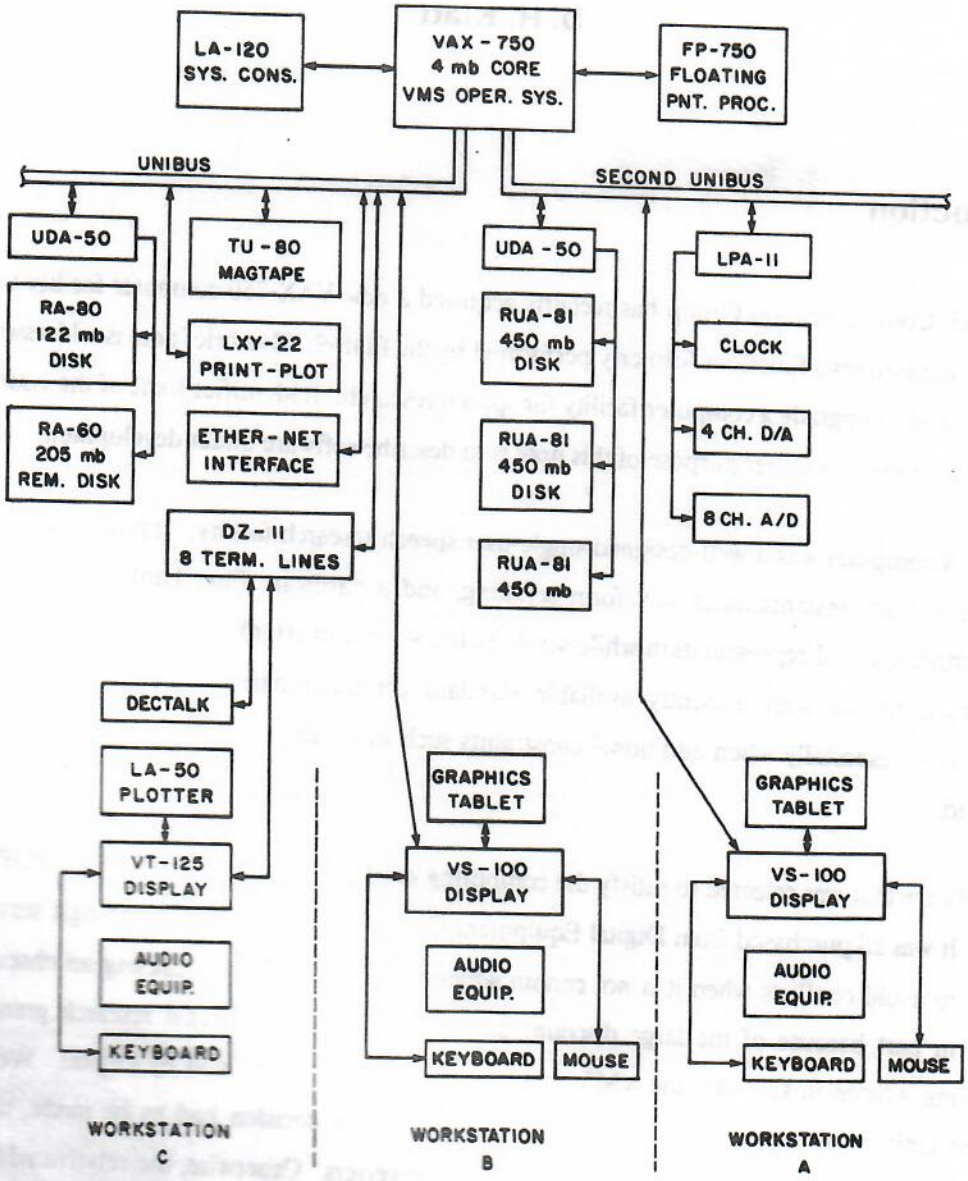

Figure 7-1: M.I.T.-SPEECHVAX hardware block diagram.

7.1.1. Hardware

The mainframe is a VAX-750 computer with 4 megabytes of core and a floating point accelerator board. With this amount of core, large programs do not produce too much swapping or page faulting activity, and two VS-100 display work stations (requiring about 1 megabyte) can be driven.

A large RUA-81 non-removable Winchester disk system provides 1350 megabytes of storage for waveforms and for the results of speech analysis activity, while a smaller 122 Mb RA-80 non-removable Winchester disk holds the system software. At 10,000 samples per second and two bytes per sample, the data disk could hold as much as 60,000 seconds, or 16 hours of speech. Modern large data bases require this kind of storage. For example, a Texas Instruments data base of 120 words spoken by 36 talkers that we are currently examining requires about 10% of the capacity of this disk.

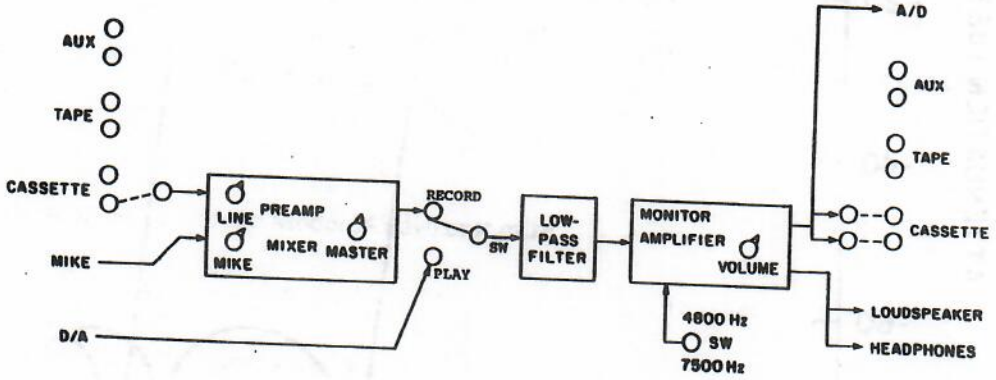

Figure 7-2: Audio patch panel layout.

Backup of system software, user software, and data is done on a relatively slow TU-80 magtape unit. In the near future, an RA-60 205 megabyte removable disk will become available for this and other uses.

Digital-to-analog and analog-to-digital conversion for three work stations (two in monaural, one work station capable of stereo) is accomplished employing an LPA-11k outfitted with a KW-11k programmable clock, an AD-11k 8-channel differential 12-bit plus-or-minus 5 volt a/d converter, and an AA-11k 4-channel 12-bit plus-or-minus 5 volt d/a converter. A software lockout mechanism prevents more than one user from accessing the LPA at a time. Had more funds been available, we probably would have obtained three Digital Sound Corporation signal conditioners instead of the LPA.

An LXY-22 printer/plotter is available for listings of programs, and for moderate quality plotting of waveforms and spectra. Examples are shown below. Work stations include an inexpensive slow LA50 plotter to permit users to save displays interactively.

The VAX communicates with other computers in the M.I.T. community (and elsewhere) by an ether-net interface to the M.I.T. Chaos network. The chaos net permits sharing of data and programs with the Lisp machines and System 20 computer managed by Victor Zue. The Chaos net is also connected to the ARPANET.

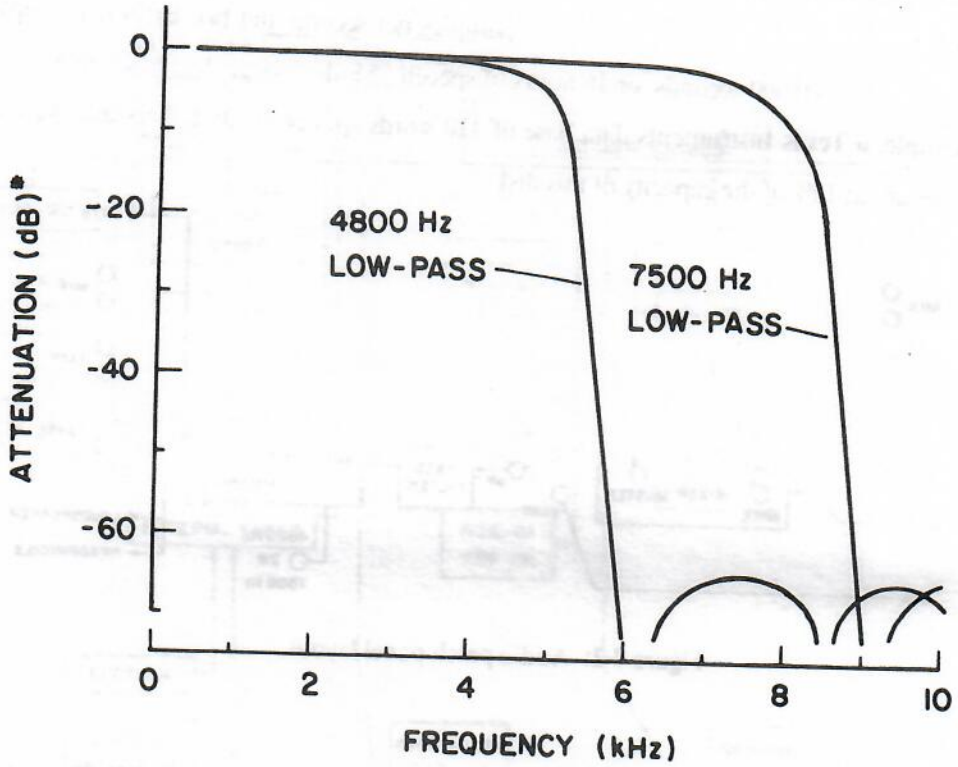

Figure 7-3: Frequency response of the 4800-Hz and 7500 Hz low-pass anti-aliasing filters inside the VAX work-station patch panel.

7.1.2. Work Stations

The three Vax work stations listed in the bottom half of Figure 7-1 consist of some or all of the components shown in Figure 7-2: a keyboard, display, mouse, and graphical tablet, as well as audio equipment including a microphone, preamplifier/mixer, cassette tape recorder, audio patch panel, monitor amplifier, speaker and earphones.

Work stations A and B use a more powerful display, a VS-100 display/graphical-device/mouse, while Work station C employs a VT-125 display. All accept the same commands from the terminal.

Any text file can be sent to a DECtalk text-to-speech box located at work station C by typing 'copy file.name dectalk'. Programs can also send text to DECtalk by simple write commands to device tta3:.

7.1.2.1. Audio Patch Panel

A specially designed patch panel and set of audio equipment is available at each work station, see Figure 7-2.

Record. The analog-to-digital conversion process is performed by the program RECORD. In order to record (digitize) a speech waveform, one must set up the patch panel to the desired configuration. The first step is to choose an audio input device from among:

- *microphone out* using a Shure Model 545 dynamic mike
- *cassette playback out* from a Yamaha Model K-1000 cassette tape recorder
- *tape playback out* from a Revox reel-to-reel tape recorder
- *aux playback out* from any device that you attach to the back of the patch panel

Choice among these inputs is governed by (1) which of the latter 3 inputs is patched to *line-in*, and (2) whether the gain controls of the Shure Model M267 mixer pass the microphone signal or the line-in signal, by turning up the gain of one or the other (or both). The patch panel *record/play* switch must be in the *record* position.

The net gain of the signal sent to the analog-to-digital converter of the Vax is determined by the master gain control of the Shure mixer. It should be adjusted so that speech peak levels read about 0 dB on the mixer VU meter (the record program indicates whether overload peak clipping has occurred during recording).

If the a/d sampling rate is set to 10,000 samples per second, the system default, then the patch panel low-pass filter select switch should be set to 4800 Hz. If the sampling rate is set to 16,000 by software commands, the patch panel switch should be set to 7500 Hz. Magnitude responses of these two TTE Model J97E 5-kohm passive low-pass anti-aliasing filters are plotted in Figure 7-3. Inexpensive high-performance fixed filters were chosen over e.g. a variable digital filter because of the large difference in cost. For sampling rates other than these, it is necessary to obtain a variable low-pass filter from the lab area and connect it at the back of the patch panel in place of the fixed filters.

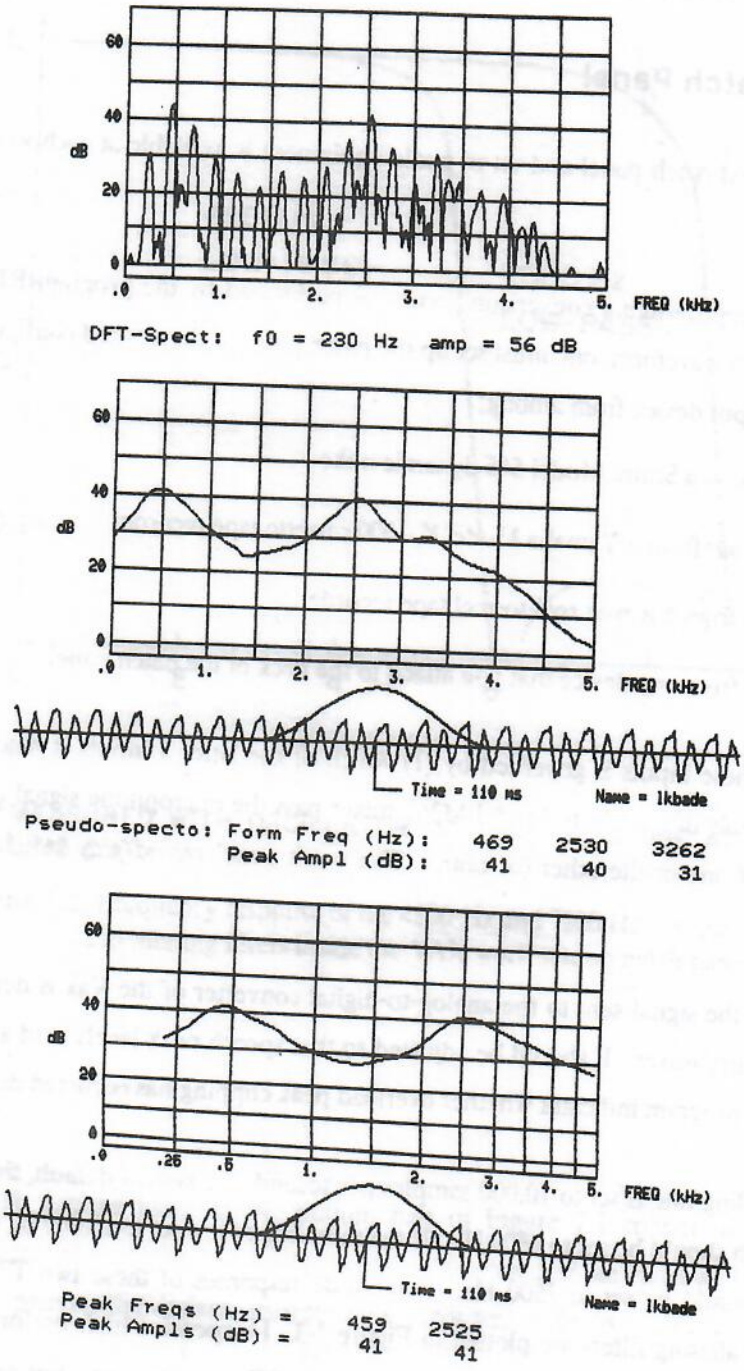

Figure 7-4: Examples of spectra produced by the KLSPEC program: dft spectrum (top), spectrogram filters (middle), and critical-band spectrum (bottom).

A modified Realistic Model SA-102 monitor amplifier drives a set of Sennheiser Model HD-430

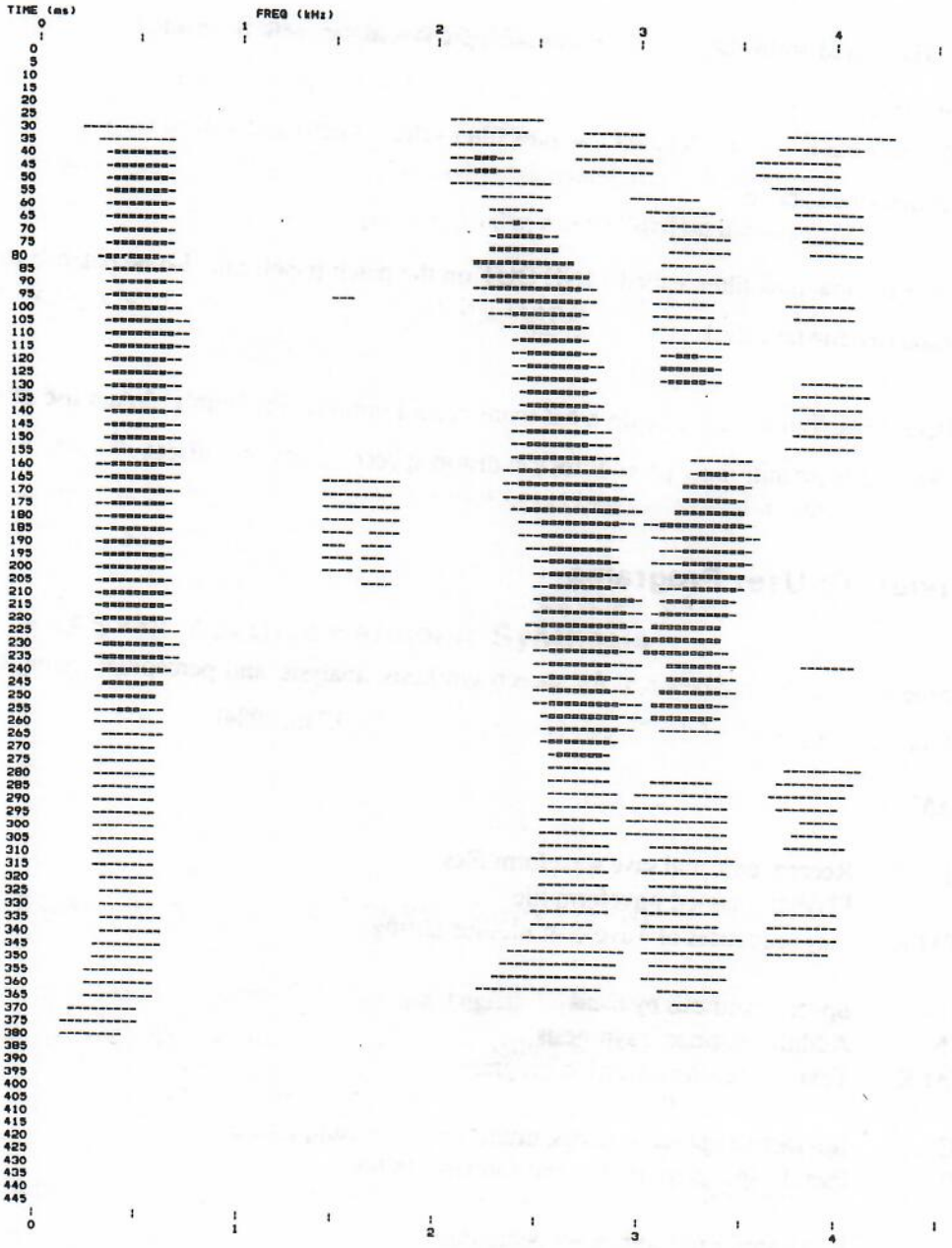

Figure 7-5: Example of a pseudo-spectrogram produced by the SPECTO program.

headphones and/or an Epicure Model 1 loudspeaker. The gain setting and tone control of the monitor amplifier have no effect on the signal sent to the a/d. The cheapest possible monitor amplifier was chosen (and the loudspeakers were a gift) — other amplifiers and loudspeakers would no doubt work as well. A separate microphone and set of good quality headphones were chosen because, on the basis of past experience, a headphone-mounted microphone does not hold up to the rough treatment found in a multiuser laboratory.

Playback. The signal from the Vax digital-to-analog converter can be monitored by setting the patch panel record/play switch to the play position. The d/a signal is low-pass filtered at 4800 Hz or 7500 Hz (depending on the position of the patch panel low-pass filter select switch) and sent to the speaker/headphone monitor amplifier described above.

The output of the low-pass filter, called LINE-OUT on the patch panel, can also be patched to the input jacks of the Yamaha cassette tape deck, reel-to-reel tape deck, or aux input line.

Cassette Deck. The Yamaha cassette deck has many record options. Preferably, do not use dbx or Dolby noise reduction because these anti-noise procedures can distort speech onsets and offsets.

7.1.3. Overview Of User Programs

A set of programs has been developed for speech synthesis, analysis, and perceptual testing that should satisfy the needs of many members of the Speech Group and visitors (Klatt, 1984).

PROGRAM	FUNCTION
RECORD	Record, edit, and save waveform files
PLAY	Playback named waveform file
MAKETAPE	Play sequences of waveform files for testing
KLSYN	Speech synthesis by hand
HARSYN	Additive harmonic synthesis
KLATTALK	Text-to-speech conversion program
KLSPEC	Interactive Speech analysis, multi waveform comparisons
SPECTO	Pseudo-spectrogram, F0, and formants to file
CFTP	File transfer to/from other computers

Table 7-1: List of user programs available on the M.I.T.-Speech-Vax.

7.1.3.1. MAKETAPE: Play Sequences of Waveform Files

This program accepts as input a list of waveform names and other parameters that constrain the automatic generation of identification tests, 4IAX tests, and single-standard paired comparison tests. The program produces an answer sheet and a subject response form that can be listed on the line printer.

7.1.3.2. KLSYN: Speech Synthesis by Hand

This program accepts user commands to draw in formant frequency data versus time, to change other constant and variable synthesizer control parameters, and then to produce a synthetic waveform file. The formant synthesizer is the same as one described by Klatt (1980) except that an optional more natural glottal source waveform can be used.

7.1.3.3. HARSYN: Additive Harmonic Synthesis

This program is similar to KLSN except speech waveforms (for voiced sounds only) are composed by adding up a set of sinusoids of appropriate frequencies, amplitudes and phases (Klatt, 1982). Unused harmonic applitudes and/or phase relationships can thereby be generated.

7.1.3.4. KLATTALK: Text-to-Speech Conversion

This program mimics the behavior of the Digital Equipment Corporation text-to-speech conversion device known as DECtalk. The program has been modified so as to produce a waveform file, and has been augmented with several debugging switches that permit observation of intermediate results, such as computed values for segment durations and synthesizer control parameter time functions. The program is thus useful in obtaining first guesses concerning appropriate synthesis values for any English word.

7.1.3.5. KLSPEC: Spectral Analysis

This program reads in up to 10 simultaneous waveform files. It displays and compares various kinds of spectra that are generated by windowing a selected portion of the waveform, computing the dft, and weighting and summing squared dft values to form 'filters'. Three spectral displays are available, the dft magnitude spectrum, a spectrogram-like filter set, and a critical-band filter set. Examples of these spectra are shown in Figure 7-4. Algorithms have been developed for estimation of formant frequency and amplitude values, and for fundamental frequency values, as illustrated in the figure.

7.1.3.6. SPECTO: Pseudo-spectrogram, F0, and Formants to File

This program takes an input waveform file and computes spectra every 5 msec. The output goes to a file that can be printed on the LXY-22 printer/plotter, and consists of a low-resolution pseudo-spectrogram display, and lists of fundamental frequency, formant frequencies, and overall amplitude versus time. One use of this display is to visualize the entire waveform prior to using KLSPEC for detailed spectral analysis. Another use is to estimate parameter values for synthesis of a duplicate of this waveform.

7.1.3.7. Other Future Programs

Future public software may include a zero-phase vocoder program (Klatt, Seneff and Zue, 1982) that permits resynthesis of a natural utterance, but with user specified fundamental frequency contour and time adjustments, inverse filtering to recover glottal waveforms from speech, and a fairly general waveform filtering program.

7.2. References

[1] Henke, W. L. (1968). Speech Computer Facility. Quarterly Progress Report No. 90, M.I.T. Research Laboratory of Electronics, 217-219.

[2] Henke, W. L. (1969). Speech and Audio Computer-Aided Examination and Analysis Facility. Quarterly Progress Report No. 95, M.I.T. Research Laboratory of Electronics, 69-73.

[3] Klatt, D. H. (1980). Software for a Cascade/Parallel Formant Synthesizer. J. Acoustic. Soc. Am. 67:971-995.

[4] Klatt, D. H. (1982). HARSYN: An Additive Harmonic Speech Synthesizer. Speech Communication Group Working Papers 1:47-60.

[5] Klatt, D. H. (1984). M.I.T.-SPEECHVAX USER'S GUIDE.[2]

[6] Klatt, D. H., Seneff, S. and Zue, V. (1982). Design Considerations for Optimizing the Intelligibility of a DFT-Based, Pitch-Excited, Critical-Band Spectrum Speech Analysis/Resynthesis System. Speech Communication Group Working Papers 1:31-46.

[2] Copies available from Speech Communication Group, Room 36-511, M.I.T., Cambridge, MA 02139, $20 each. Check should be made payable to Speech Communication Group.