A SPEECH DATA BASE AT THE UNITED STATES AIR FORCE ACADEMY

Lt Colonel Michael F. Guyote, Captain Keith A. Lewis, and Mr. Donald Lijana

United States Air Force Academy
Colorado Springs, Colorado 80840-5851

ABSTRACT

The Department of Electrical Engineering at the United States Air Force Academy is presently building a speech data base which will consist of selected speech groupings from 1000 of the Academy faculty and students. Data collection is automatic and supervised by computer programs which request information on speaker background, accents, age,etc. and run the actual data collection procedures. The information on individual speakers is placed onto a computer permanent file and is also written onto one channel of the data base tape. Data is encoded onto four channels of Beta format videotape. The data consists of: Pulse code modulated (PCM) speech from a high-quality capacitor microphone, PCM speech from a standard United States Air Force close talk microphone, analog speech from an electroglottograph, and encoded speaker background information.

INTRODUCTION

The Department of Electrical Engineering at the United States Air Force Academy has installed a state-of-the-art speech analysis facility. This paper describes the collection system actually installed to implement the proposal discussed in "Development of a Speech Research Capability at the U.S. Air Force Academy" by Stanton and Burge [1]. The facility consists of soundproof booth, extensive audio hardware, and both purchased and self-developed analysis software. This facility will give Academy faculty expanded opportunities for computer analysis of speech patterns, particularly in continuous speech. The research project is funded by the Rome Air Development Center, USAF Systems Command, Griffiss AFB, New York. The long range project goal is to develop usable algorithms which would allow aircraft computer systems to recognize aircrew speech commands. The short-range goal is to develop a large speech data base which can be used by the USAF Academy and other organizations as suitable test data for ongoing speech recognition projects.

Since the goal of USAF speech recognition research is to establish a viable machine-based speech recognition system for use in operational environments, an ideal data base should be representative of those individuals who will be operating USAF aircraft now and in the future. The faculty and students at the Air Force Academy comprise such a group from which to compile this data base. The Academy Cadet Wing consists of over 4400 students (approximately 4000 male, 400 female) of age 18 to 24. These students come from the 50 states, US territories, and allied countries. Nearly 70% of these cadets will go into undergraduate pilot training following graduation. After obtaining their pilot ratings, they may be assigned to any aircraft in the USAF inventory. The speech patterns collected from this group will comprise an ideal data base for use with machine-based speech recognition systems to be used in future aircraft. The members of the freshman class in particular retain most of the salient characteristics of their regional dialects. This provides an additional degree of difficulty to speech recognition systems, which must be capable of responding to a wide range of accents, dialects, etc. if they are to be used for anything other than the most simple of tasks.

DATA BASE SPEECH

The speech data base consists of the phonetic alphabet, the digits zero through nine, and selected sentences designed to provide a wide range of phonemic structure. During data collection, each speaker is asked to read the elements of the data listed above. A total of seven sentence library stored in system memory. Each individual is requested to repeat each element and sentence of the data base three times. These repetitions are not in succession; for example, the phonetic word "kilo" will not be requested in the same order during each portion of a specific data collection run. This is done so as to provide lessened possibility of a particular data word affecting subsequent words.

Audio data is taken from three sources: A high quality capacitor microphone, a close talk, noise-canceling pilot microphone, and an electroglottograph. The high quality microphone is located approximately 4 cm from the speaker's lips and provides the highest fidelity voice signal in the data base. This is intended to be a reference signal against which the other data can be measured. The capacitor microphone output is amplified and the output is fed to one of two PCM inputs. The pilot's close talk microphone is located approximately 0.5 cm from the speakers lips. The close-talk microphone output is amplified and fed to the other PCM input. The close-talk amplification system introduces some broadband noise into the final output and is intended to simulate actual voice output from an aircraft intercom system. PCM outputs are fed to the video input of a Beta format videorecorder.

The glottograph output is taken from the speaker's larynx. Laryngal output is measured as a function of changing capacitance of the sensors, which are placed on either side of the test subject's larynx. The output yields data on the pitch or excitation frequency, of the subject's larynx, but little useful data on the actual speech.

All speech data base records are written onto Beta format videotape. The two voice outputs (high quality and pilot's close talk microphone) are encoded into PCM format and written onto the video channel of the video recorder. Digital encoding is done by a Sony PCM-701ES Audio Processor. The glottograph output is written onto one of two audio channels contained on the video system. Encoded information concerning the speech data is written onto the other audio channel. This information is encoded via a standard frequency shift modem operating at 300 baud.

Due to the extensive error correcting abilities of the PCM decoding unit, both channels of the encoded data written onto the video channel are capable of being duplicated many times without noticable degradation. The PCM system was chosen for precisely this reason to allow extensive duplication of the speech data for use by those who request it. The audio channels contain data with an extremely limited bandwidth and as such can undergo multiple copying without degrading relevant information. The system layout is shown in Figure one.

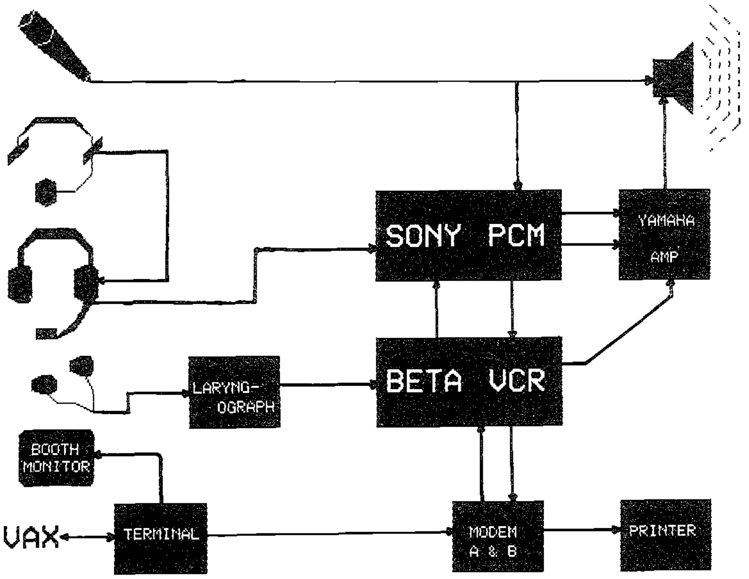


Figure 1. Speech Data Collection System Layout

AUTOMATED DATA COLLECTION PROGRAM

Collection of such a large amount of data requires standardization as well as an efficient method of gathering the speech inputs. The data base collection system is controlled by a set of Fortran programs run on a Digital Equipment Corporation VAX 11/780 computer. Program operation requires an operator and one subject. The subject is placed in a soundproof booth. The two microphones and the glottograph are adjusted to fit the subject. The operator calls up the data collection program, and the program prompts both the operator and the subject as required.

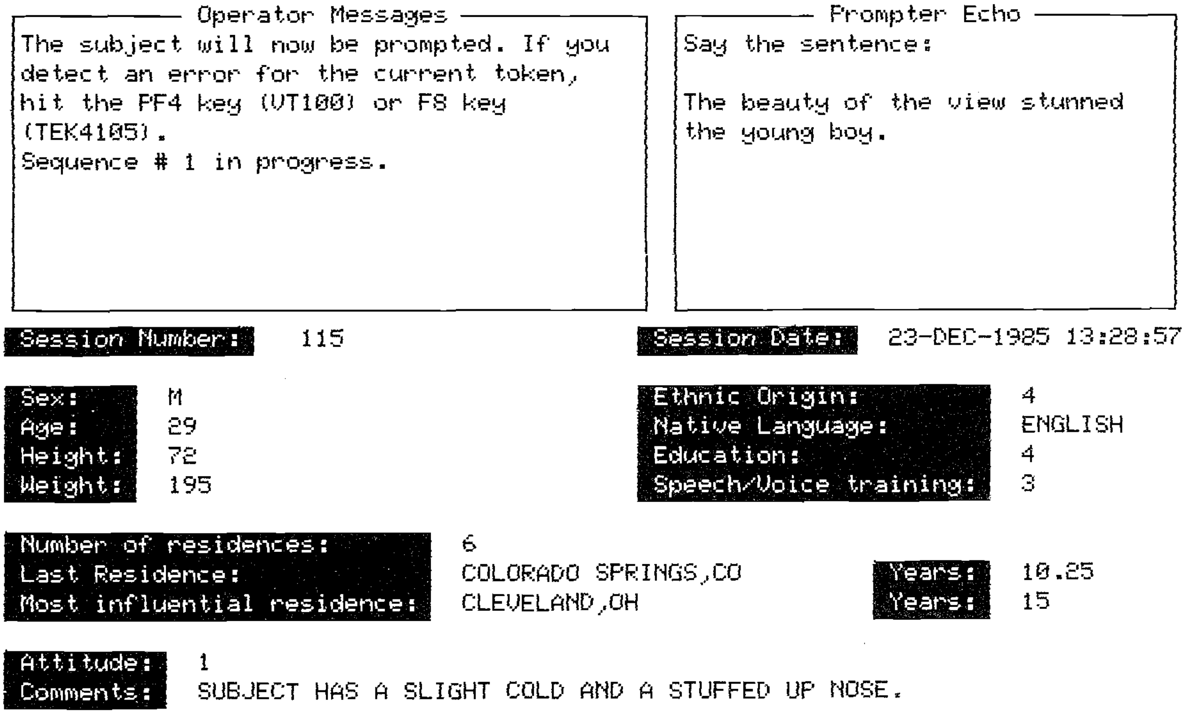
At program start, the data collection program first requests background data on the test subject. This data includes: Session number and date, subject age, sex, height, and weight, subject's ethnic origin, native language, education, and whether subjects have had speech/voice training. The program also requests information on factors which may influence the subject's speech. This information includes locations and time spent at each location during the subject's early life. An example of the screen prompt given to the data collection operator is shown in Figure 2. The operator screen is divided into three portions: two small rectangles containing operator and subject prompts, respectively and the remainder of the screen containing the requested subject background data.

Actual data collection is semi-automatic. After the subject has been fitted with the microphones and glottograph and the program has gathered the requested background data, the program prompts both subject and operator through individual CRTs. The subject sees only the prompts relevant to his/her speech data. The operator has a headset which allows monitoring of subject speech. Should the subject mispronounce a requested word or sentence, the operator will place a mark on the data base (by pressing a selected key on his console). This mark will alert subsequent data analysis routines of the suspect nature of that particular section of the speech data.

The collection program also writes the requested data text to a printer which is in the speech laboratory. This data text is annotated with the session number in order to provide a reference as to what was spoken on any particular speech session. A standard telephone modem (300 baud) is also attached to the printer input lines. The printer speech data output is fed through this modem and is written onto one of the low quality audio lines as an information track. The purpose of the information track is to provide for future closed-loop automated analysis programs. These programs could access any of the three audio data tracks, perform an analysis on the speech data, and use the fourth data track to compare the analyzed speech with the code for the actual speech. Should the subject's speech be defective in some way, the operator initiated mark mentioned in the previous paragraph will be noted by the analysis program. A change which will be implemented in the near future will allow the complete subject information package to be encoded onto the information track. This will be located at the beginning of the speech session and will allow analysis systems to have this additional information for use.

REFERENCE

[1] Stanton, Major Bill J. Jr. and Burge, Major Legand L. Jr., "Development of a Speech Research Capability at the U.S. Air Force Academy" IEEE Region 5 Conference, Lubbock, Texas, March 1985.



```
------------ Operator Messages ------------    -------- Prompter Echo --------
The subject will now be prompted. If you       Say the sentence:
detect an error for the current token,
hit the PF4 key (VT100) or F8 key              The beauty of the view stunned
(TEK4105).                                     the young boy.
Sequence # 1 in progress.

Session Number:   115                          Session Date:  23-DEC-1985 13:28:57

Sex:      M                                    Ethnic Origin:            4
Age:      29                                   Native Language:          ENGLISH
Height:   72                                   Education:                4
Weight:   195                                  Speech/Voice training:    3

Number of residences:        6
Last Residence:              COLORADO SPRINGS,CO          Years:   10.25
Most influential residence:  CLEVELAND,OH                 Years:   15

Attitude:   1
Comments:   SUBJECT HAS A SLIGHT COLD AND A STUFFED UP NOSE.
```

Figure 2. Operator's CRT Presentation