[redacted] D15

N93-72627

SPEECH SYSTEMS RESEARCH AT TEXAS INSTRUMENTS

DR. GEORGE R. DODDINGTON

TEXAS INSTRUMENTS INCORPORATED
DALLAS, TEXAS

515-32
176351

I. TI Capabilities

Texas Instruments supports a Speech Systems Research branch in its Central Research Laboratories. The charter of this branch is to foster the development of new TI business opportunities through development and application of automatic speech processing technology. Seven speech research programs are currently active: Two corporate funded programs determine the strategic direction of our speech research; these are programs to develop automatic dictation technology and low-cost vocoder technology. Three programs are externally funded by Rome Air Development Center to develop and apply advanced but near-term speech processing technology. These programs are: "total voice" speaker verification, limited vocabulary continuous word recognition, and automatic language identification. Finally, two programs are supported internally by TI's operating divisions.

The Speech Systems Research Computer Laboratory contains a variety of computer systems for speech research, system evaluation and product development. System 1, the principal research system, is diagrammed in Figure 1. The salient features of this system include real-time speech I/O, a 500 Mbit disk, a Floating Point Systems AP120B array processor, and a Tektronix interactive graphics terminal with hard-copy.

II. TI Achievements

A. Voice Authentication

Although Texas Instruments has been active in a variety of speech processing problems including speech analysis, speech synthesis, word recognition and speaker verification, most of our research effort in the past has been devoted to the development of speaker verification technology. Speaker verification, in its operational format, we refer to as voice authentication. A sequence of 3 programs, funded by RADC and beginning in 1972, led to the development of a voice authentication technology capable of meeting

Figure 1. Functional Diagram of the TI Speech Research Computer System #1

BISS[1] performance requirements of less than 1% user rejection at less than 2% impostor acceptance. In retrospect, there were 3 primary problems that we solved in this development:

1) Enrollment

   Enrollment of a user on the voice authentication system must be performed in a single session. Enrollment is difficult for the following reason: Speech data collected in a single session is relatively self-consistent but not representative of an ensemble average over many sessions. Therefore, the initial reference data is biased and the initial estimate of speaker variance is invalid. This problem has been effectively solved by requiring extra speech input data in a special 4-session "post enrollment" strategy and, recently, an additional special 8-session "post-Post enrollment" strategy.

2) "Goats"

   Voice authentication system users are classified as either "sheep" or "goats". The sheep are well behaved and far outnumber the goats. The system performs well on sheep. The speech data of goats exhibit high variance, but the voice authentication system must perform well for everyone. Uniform performance is achieved by a carefully designed decision function which requires more speech data from the goats while, at the same time, not prejudicing the verification decision against them.

3) Discipline

   Voice authentication system users have little interest and little interaction with the authentication system. Authentication utterances often have false starts or are imbedded in extraneous speech data. The verification system must be able to extract the proper input data and discriminate between proper data and garbage. Time registration

---

[1]Base and Installation Security System, a program administered by the Air Force to define and develop future military security systems.

through energy end-points cannot solve this problem. Proper time registration is achieved through a continuous spectral matching algorithm.

Texas Instruments has controlled access to its Corporate Computer Center by voice authentication for the past three years. This system is in operation 24 hours/day and has provided over ½ million verifications. Current system performance is ½% user rejection at an impostor acceptance level of 1½%. Most user rejections are attributable to noncooperative quirks of the user.

B. Word Recognition

Texas Instruments has operational real-time demonstrations of word recognition for isolated and connected words, enrolled and independent speakers, and small vocabularies up to 50 words. Large vocabulary recognition has been performed in non real time for vocabularies of greater than 1,000 words. Also, the development of automatic language identification has been sponsored by RADC. One significant result in language identification is the demonstration of improved identification by normalization of long-term spectral averages. Although long-term spectral averages have been shown to be useful in discriminating languages, the wide variety of recording conditions encountered in our data base invalidate such utility. With the incorporation of spectral normalization, results of 5-language identification task have been improved to 80% correct on excerpts of two minutes duration.

"Total Voice" speaker verification involves two speech processing tasks: speaker independent recognition of an identifying sequence of six spoken digits; and speaker verification using the user identification and the same identifying speech input data. The application requires speaker independent recognition of a connected sequence of six digits with less than 1% sequence recognition error. These severe application requirements have been achieved by incorporating two check digits in the sequence for improved recognition accuracy and by "forbidding" certain digit pairs such as "three-eight".

III. Fundamental Problems

I have ordered below four classes of problems that must be faced in the development and deployment of an automatic speech processing system:

A. Speech Science

The lack of speech science limits the capabilities of speech processing systems. So, how do we go about getting speech

science? Careful direction is exceedingly important because one can easily drown in the vast, uncharted oceans of speech phenomena. In my opinion, a very good way to get speech science is to identify an important real application for speech processing and to persevere in developing speech technology for this application. This approach, which I refer to as the "correct" approach, is contrasted with the traditionally popular method in Figure 2.

B. Cost

Cost is a very important consideration in automatic speech processing for two reasons: First, speech processing is a complex problem which is inherently costly, at least in terms of computing power. Second, system cost effectiveness is usually measured in terms of the efficiency of a person, at least in the case of speech recognition. Cost/benefit tradeoffs must be carefully made between speech and other alternate media.

C. Performance Forecast

It would be nice to do an experiment in the laboratory and be able to say with confidence that the laboratory results will be realized in the operational system. This rarely, if ever, happens because the laboratory data is not representative of operational data. Sometimes the discrepancy between laboratory results and operational results is embarrassingly large. One important reason contributing to this is the typical favorable mix of sheep with goats in laboratory experiments. (Speech researchers are usually "super sheep".) System performance depends very strongly on this mix. A large number of subjects is required to properly evaluate system performance. Figure 3 is included to provide some perspective on the sheep/goat problem. This figure shows a histogram of user data variance for an operational voice authentication system.

How much data must be included in a laboratory experiment to provide a good performance forecast? My rule of thumb answer to this question is that enough data must be collected to provide 30 errors. Assuming that each trial is independent, thirty errors will provide you with an estimate of the true error rate within ±30%, for the given data context, with 90% confidence. Suppose for example that you anticipate 1% error for a certain word recognition system. This implies that at least 3,000 spoken words must be collected to provide the desired confidence interval on error rate. Note however, that the trials must be statistically independent. Will you collect 3,000 words from one speaker or one word from each of 3,000 speakers?

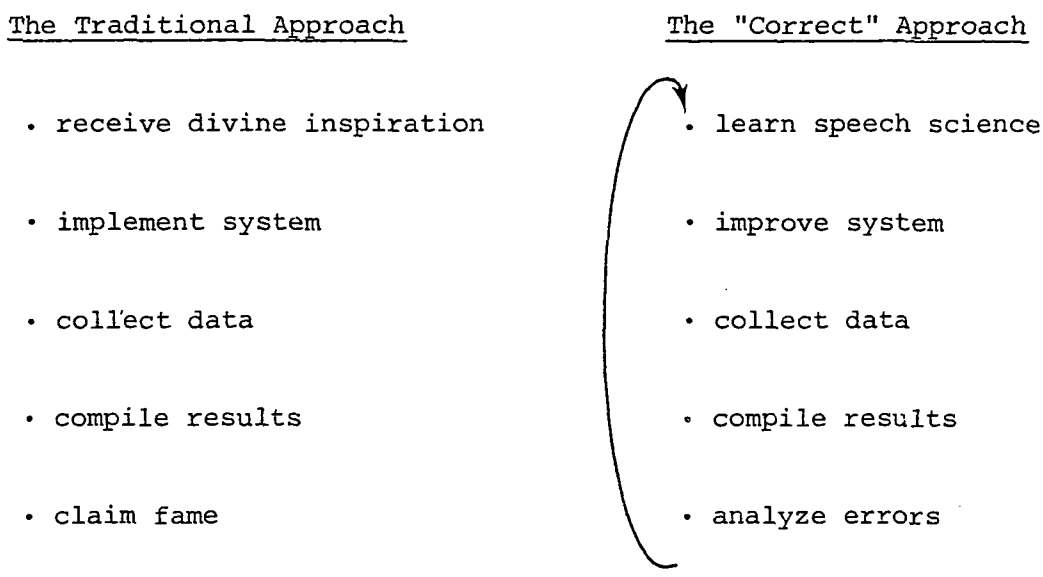

Figure 2. Basic Steps in Developing Speech Technology

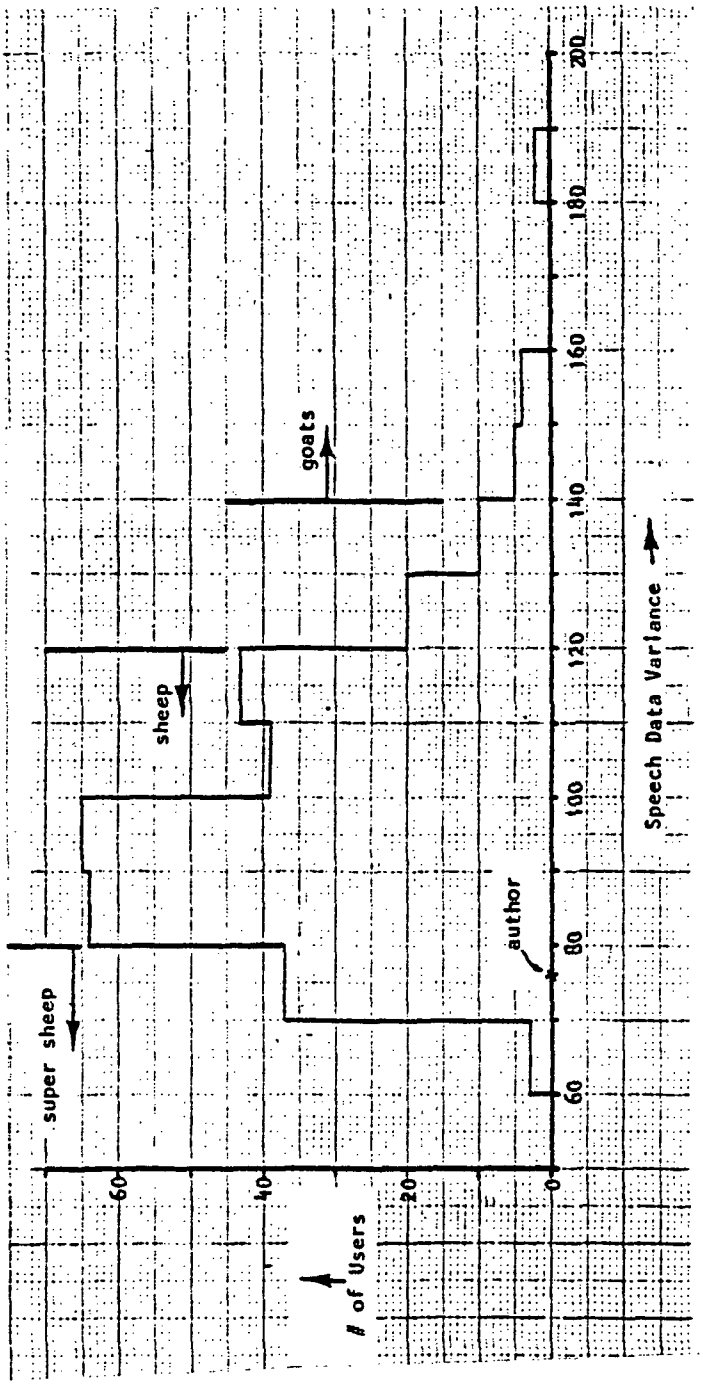

Figure 3. Histogram of User Data Variance for Operational Voice Authentication

D. The Human Factor

User acceptance is a critical factor in speech processing systems. For word recognition this includes not only recognition performance, but also other recognition characteristics. For current technology, isolated word recognition machines, a most important operating characteristic is the requirement for pausing between words. This is a nontrivial skill to perform reliably and is a major underlying factor in initial performance degradation. Fortunately, system performance is often aided through the adaptation of the user to the system. This includes learning to speak clearly and loudly to the system in spite of the fact that the microphone is often less than 1 inch from the mouth. Loud, clear input stabilizes the speech data and improves system performance. Such user adaptation is clearly demonstrated in Figure 4. Figure 4 is a plot of user data variance as a function of session number for an operational voice authentication system. The subtle feedback in this system has provided a user learning time constant of 2,000 sessions.

IV. Technology Forecast

Progress in speech systems development is tied closely to developments in computer technology. Advanced speech system capabilities will require inexpensive yet highly competent high speed data processing. The speech processing system will comprise a general purpose CPU with speech input through a special purpose speech preprocessor and feature extractor. An important cost element is this speech preprocessor unit. TI is currently developing, under contract with ARPA, a one-chip speech analyzer using CCD technology. This chip implements a 19-channel sampled data filter bank with on-chip 4-bit A/D conversion and multiplexing.

Low cost speech preprocessor technology coupled with advances in microprocessor performance is anticipated to have substantial impact on speech system competence, cost and market size by 1980-1982. At this time useful and affordable capabilities will be introduced for connected word recognition and narrowband digital voice communication.

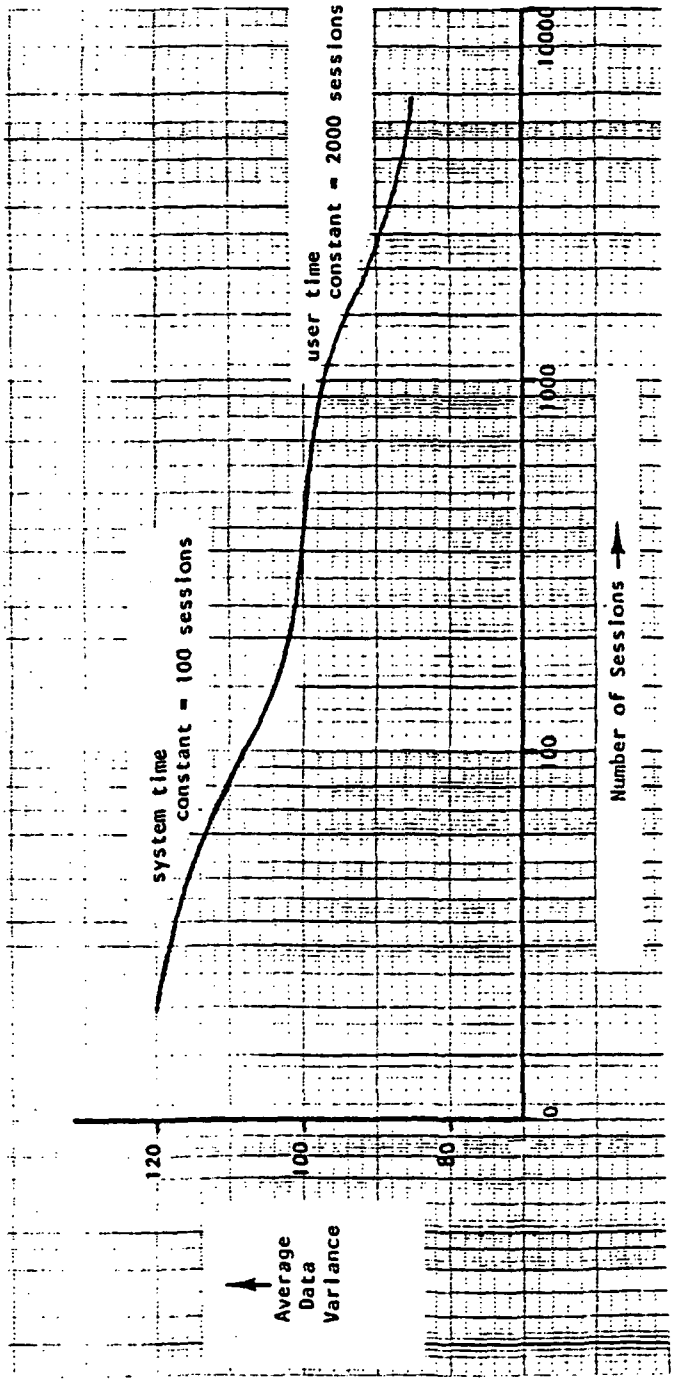

Figure 4. User Data Variance as a Function of Session Number for Operational Voice Authentication

BIOGRAPHICAL SKETCH

George R. Doddington

Ph.D. in Electrical Engineering, University of Wisconsin
M.S. in Electrical Engineering, University of Wisconsin
B.S. in Electrical Engineering, University of Florida
Professional Engineer, State of Wisconsin

At Texas Instruments, Dr. Doddington has directed programs of speech research encompassing advanced speech processing techniques. This work has included interactive simulations of word recognition, speech segmentation and analysis, and speaker verification. Dr. Doddington's doctoral study emphasized communication theory, control theory, probability theory, and neurophysiology. His doctoral research was conducted at Bell Telephone Laboratories during 1969 and 1970. This work comprised the development of a method of nonlinear time normalization of speech and the implementation of this method in a system of speaker verification. Dr. Doddington joined Texas Instruments in 1970. Dr. Doddington's master's thesis comprises a generalized theory for relating the gross operating characteristics of chromatographic systems to the statistics of molecular behavior. Dr. Doddington was employed at the Federal Communications Commission from 1960 through 1963, during which time he designed and developed a secrecy coding system for radio-teletype communication. In 1964 he received the bachelor's degree with high honors for his thesis comprising the theory and practical implementation of a method of linear amplification approaching 100 percent efficiency.

DISCUSSION

Dr. George Doddington

Q: Mark Medress, Sperry Univac: Do you have any feeling for how much action you got out of the check digits in your total voice verification system?

A: Okay, I didn't talk about the performance of that system, really, except I did say we had eight errors in about a thousand trials. Yes, I have some feeling for that and I'll tell you. The eight errors in a thousand trials represented about one percent error. Now that's on the six digit sequences themselves. We have run some experiments using the digit recognizer component, in the sequence recognition strategy, and we've gotten about 95 percent correct digit recognition. Now those two performance figures are not comparable. The digit recognition is for digits, 95 percent correct, and the sequence recognition is for sequences, and that's about 1 percent error.

Q: Mark Medress: That's 95 percent of the words in those sequences are correct, is that what you're saying?

A: That's right. 95 percent of all digits were correct.