D10

[redacted]

N93-72622

50-32
176346

MILITARY APPLICATIONS OF AUTOMATIC SPEECH RECOGNITION & FUTURE REQUIREMENTS

DR. BRUNO BEEK
EDWARD J. CUPPLES

ROME AIR DEVELOPMENT CENTER
GRIFFIS AFB, ROME, NEW YORK

The objective of this paper is to provide an updated summary of the state-of-the-art of Automatic Speech Recognition (ASR) and its relevance to military applications. This paper follows the format of the recent IEEE paper on "An Assessment of the Technology of Automatic Speech Recognition for Military Applications (I)". Until recently, speech recognition has had its widest application in the development of vocoders for narrowband speech communications. Presently, development in ASR has been accelerated for military tasks of Command and Control, Secured Voice Systems, Surveillance of Communication Channels and others. Research in voice control technology and Message Monitoring Systems and Automatic Speaker Verification Systems are of special interest. Much of the emphasis of today's military supported research is to reduce to practice the current state of knowledge of ASR as well as directing research in such a way as to have future military relevance.

Already in use are voice data entry systems for a number of interactive command and control functions. These applications are limited to a small vocabulary, speaker dependent, isolated word recognition system. These highly reliable systems allow for hands-free source data entry of the digits and a limited set of control words and phrases. As such, the voice data entry system eliminates manual transcription and keying operations in hand/eyes busy mode of operation.

A large number of potential systems for military applications are now under development. These include:

1. Digital Narrowband Communication Systems: A large research and development program has yielded a number of digital communication systems mainly using linear predictive encoding analysis. These various systems have been analyzed and equipments are now being developed for operational tests. These efforts have resulted in the need for the fabrication of a low cost speech input/output front end. Major research in this area is underway.

2. Automatic Speech Verification: An advanced development model has been fabricated for secure access control applications and has been tested under laboratory and operational conditions. These tests demonstrated that speaker verification is a viable technique.

[redacted]

3. On-Line Cartographic Processing System: Studies are under way to use speech recognition and voice response techniques with cartographic point and trace processing systems. Equipments have been fabricated and tested under operational conditions. Other research and development programs are now undergoing test and evaluation.

4. Word Recognition for Militarized Tactical Data Systems: Word recognition, speaker verification and voice response will be used for message entry to a tactical data system. Equipment has been fabricated and is being or has been tested.

5. Voice Recognition and Synthesis for Aircraft Cockpit: Existing word recognition systems are being tested and evaluated under simulated cockpit environments.

These system studies and implementation are in various stages of investigation and/or development and represent a practical utilization of the emerging speech recognition technology.

Numerous independent and integrated efforts have been undertaken to further develop the speech generation, reception and reproduction phenomena for specific military and civilian applications. This has increased the interaction among scientists in fields of acoustic phonetics, linguistics, signal processing, computer science, etc. Fortunate or unfortunate, the major emphasis on present day funding is dealing with the practical application with existing word recognition systems with only minimal consideration to some major improvements which are necessary before these techniques can be used in ever increasing military systems.

1. Automatic Message Monitoring: As the technology advances, the utilization of message monitoring functions shall find increased relevance for command and control functions.

a. Keyword Classification: The goal of keyword recognition/classification is to recognize a keyword or a set of keywords embedded in narrow-bandwidth conversational speech as expected from a radio link. The reconnaissance of large amounts of speech information requires a need for economical data editing and scanning. An automatic method of detecting and classifying keywords would perform this function. The difficulties in providing this speech processing function are that the speaker is unknown, the speech is continuous and the speech signal is often degraded by noise and distortion. Problems also exist due to coarticulation and context since the acoustic representation can be affected significantly by the acoustic environment of the keyword.

The technique which seems to have had the greatest amount of success is to recognize acoustic events simultaneously in free running speech. A sequential logic, made up of acoustic events, can be designed for a keyword. This approach postulates that keyword detection will take place when the needed sequence of acoustic events occur. Results to date indicate that detection rates are approximately 85% with 10-15 false alarms per hour.

b. Language Identification: Numerous techniques have been applied to Language Identification due to a fluctuating military interest in this problem. The most general approach has been, as in speaker identification, to use pitch and spectral features. Recently, emphasis has been placed on the recognition of acoustic events within the speech signal which are reliable, stable, and speaker independent. This work, and the work of devising methods of converting the acoustic signal into a chain of linguistic elements, offers considerable promise for a solution of this problem.

Experiments have been performed with perfect chains of this type, formed from a phonetician's hand transcription of speech in various languages. These experiments show that the statistics of these chains, especially higher-order statistics of digraphs and trigraphs, are a very powerful means of discriminating between languages given one to two minutes of speech.

Experiments were also performed on five languages involving over a hundred different speakers using acoustic events extracted from continuous speech. The data on each speaker was collected over a period of time to determine the effects of speaker variabilities on recognition performance. Fairly encouraging language recognition scores for all five of the languages has been attained. Continued research and development is necessary in order to implement an on-line, real-time operational system.

2. Command and Control: Command and control, in the context of this paper, means voice communications with machines. These include:

Limited Word Sets

Connected Word Recognition (Limited set of words)

Continuous Speech Recognition/Understanding

a. Isolated Word Recognition: Speaker dependent isolated word recognition for all intents and purposes has been solved under laboratory conditions, but engineering problems still exist. Even with these problems a number of commercial companies in the United States and England are marketing isolated word recognizers. Some voice date entry systems are already operating in a variety of industrial applications. These include:

1) Automatic Sorting Systems, distribution of parcels, containers and baggage.

2) Voice Programming for Machine Tools

3) Inspection System, e.g., measurement of TV face plates

4) Automobile quality control inspection

5) Various military applications.

Generally, two types of problems exist, those dealing with the speech processing/recognition process and those related to the incorporation of the word recognition device into an operational system. The first set of problems is:

1) Development of low-cost, light weight human acceptable, noise cancelling microphone.

2) Training various individuals to use the word recognizer. This includes minimizing the number of repetitions per word. This problem becomes more severe as the vocabulary increases.

3) Development of an adaptive technique to upgrade the word patterns as a function of time.

4) Decrease the error rate. Even a 1-3% error rate can cause large time delays in data input.

5) Others

System problems include the development of a gracefully interactive system that an individual can feel comfortable with. This includes:

1) A learning procedure. Teaching the individual to use the system in such a way that he uses it in a normal everyday manner.

2) Develop error correcting methods to limit backup time.

3) Development of optimum system feedback (visual, audio or both).

4) Others

b. Multi-Speaker, Isolated Words: Emphasis is being placed today on the development of an isolated word recognition speaker independent system which operates over a good quality telephone network (bandwidth 300-3500 Hz). Experiments have shown that a relationship exists between word size and speaker dependence as a function of recognition accuracy. Although most multi-speaker systems tend to be insensitive variations in the rate of speech, problems still exist due to variations in a talker's speech characteristic and talker dependent traits. Recent experiments have shown a multi-speaker recognition system to be highly accurate in recognizing a small set of words (10-20) even in the context of connected speech. However, a number of constraints had to be introduced. These include:

1) Knowledge of the number of words in a string

2) Fixed number of digits with error correcting codes

3) Requirement of a short learning phrase.

c. Continuous Speech Recognition (Speech Understanding): The Department of Advance Research Projects Agency has recenlty completed a program to recognize sentences that can be produced from a known vocabulary (Lexicon), with known subject matter (Semantics), using known grammatical rules (Syntax). The results of this program are now being published. In addition, a study is underway to determine the major scientific contributions of the work. The ARPA project considered the task domains as data retrieval.

3. Speech Enhancement: Programs are underway to enhance the intelligibility of speech signals transmitted over a low-quality communication channel. Techniques under investigation offer potential for the development of an automatic system for attenuating non-speech signals which accompany speech and interfere with and occasionally obscure the information bearing parameters of speech.

Two techniques were developed for enhancing the S/N ratio of speech received over a noisy channel. The first method is intended for use when the noise is wideband and random. It is similar to homomorphic filtering, however, the spectrum rooted (rather than logged) before being transformed. Limited test results showed the effective S/N ratio increased somewhat without seriously distorting the character of the speech.

The second enhancement technique is useful when the interference consists of tones or can be decomposed into tones. It consists simply of transforming the speech-plus-noise to the spectrum domain, detecting and attentuating the tones, and retransforming the enhanced spectrum to the time domain. By use of this method, speech signals that were barely detectable at S/N ratios below -26 dB were made fully intelligible.

The enhancement processing in its present configuration requires extensive hardware and software for real-time operation. The reason for this complex system is to transform the noisy speech signal in such a way as to easily remove the noise portion without appreciably affecting the voice signal.

Systems of this type are in various stages of exploratory research and advanced development. In the near future we shall see the operational implementation of these systems.

4. Security Applications: Automatic speech recognition may soon be extensively used in the area of security. First and foremost in terms of military applications is the problem of automatic speaker verification/identification. This security problem and others are in different stages of solution.

a. Automatic Speaker Verification (ASV): Speaker verification means the verification or rejection of an individual based on his speech patterns. In general, each individual known to the ASV System has on file a number of samples of his speech. When he wishes to be verified he must first identify himself to the ASV system via a badge reader, keyboard, magnetic card, etc. Once he has identified himself, the ASV system requests that he speak a number of sentences, phrases or words. After the individual complies, the ASV system analyzes the incoming data, compares it with its reference file and makes a decision either to accept or reject the individual or request additional data. ASV technology has a number of advantages not always available in other speech recognition technologies. Namely, the speaker is cooperative and is attempting to gain access to some function and hence will be on his best behavior. The speech data spoken by the individual is known to the ASV system; sentences and words are chosen to provide the greatest amount of discrimination. The acoustic environment can be either controlled (good signal-to-noise ratio) or noise cancelling microphones can be used. Analysis of an individual's reference speech data may lead to extraction of customized features for that individual to maximize speaker discrimination. In the operational mode, where the individual tests the system, the ASV can, by analyzing the speech and finding it deficient (not loud enough, garbled, etc.), can request individuals to repeat. Further, the communication channel can easily be made identical for both reference and test.

A number of techniques are being investigated by Bell Telephone Laboratories, North American Rockwell and others. Perhaps the most successful is that of Texas Instruments (supported by the United States Air Force/RADC).

This technique was designed to have less than 1% rejection of true speaker, and less than 2% acceptance of impostors.

The ASV system has been tested under the conditions of mimicry, day-to-day speaker variability, colds, sinus congestion, and respiratory ailments. In general, the technology can handle most operational requirements and achieve low Type I errors (true speaker rejections) as well as Type II errors (impostor acceptance) with certain tradeoffs depending on the specific application. For example, Type II errors can be lowered at the expense of increasing Type I errors and vice versa. Both types of errors can be reduced at the expense of having the user utter more speech data.

The technology has been tested operationally by the Air Force Electronic Systems Division (ESD/BISS) and Mitre Corp using a large group of military personnel. In addition the system has been in day to day use at the Texas Instruments Corporate Information Center for three years and has achieved a Type I error rate of 0.5% and 1.6% Type II error rate.

This is a research area of interest to the military (e.g., accepting or rejecting front-line tactical reports), in industry (e.g., controlling access to restricted areas), and in commerce (e.g., controlling access to money or information). There is a stated military requirement (ROC) for a device to handle this problem under field conditions. The technology is presently being advanced via the development of systems which operate over phone lines and systems which require an individual to only speak a 6 digit code for identification and verification.

b. Speaker Identification (Recognition): This section describes several problems dealing with automatic speaker identification of an individual based on his voice characteristics. Speaker identification is discussed in the context that recognition decisions are rendered automatically from continuous speech uncontrolled as to context. Hence, the speaker's reference library and testing unknown speech samples can be generated independent of spoken text. This advantage is important when one is collecting and analyzing speech data from an uncooperative speaker. At the present time, applications of greatest interest are the use of the voice signal for personnel identification for access control, monitoring communications channels and computer security. Conditions that make this problem difficult in practice or real application are:

1) the communication system is of poor quality

2) speakers may be non-cooperative

3) recording and/or channel conditions are different for reference and test samples

4) deciding whether the speaker is a member of an original group of speakers

This kind of problem arises in the military when, for example, one tries to keep track of a unit that is communicating by radio. Presently, emphasis is being placed on analyzing an individual's speech signal when particular phonetic events occur. In this manner, the vocal tract transfer function can be determined by a detailed spectrum analysis. Combining a number of these measurements for a number of phonetic events can provide the data to identify a speaker.

A number of speech investigations have studied the use of linear least-square, inverse filter formulation to estimate the format trajectories of selected phonetic events.

Summary and Conclusions

Although significant progress has been made in all areas of automatic speech processing in the past ten years, we still have a long way to go before they can be incorporated into the military inventory.

We have attempted to describe the various military problems, existing solutions and how they lead to present and future technology requirements. The utilization of these automatic speech processing systems have lead to new emphasis in technology development, namely, the requirements for a low cost speech processing front end system to be used for automatic word recognition, speaker identification and vocoder applications. Consideration is being given, by various systems choices, to the "best" methods of incorporating word recognition technology as a peripheral device in their system specifications. Numerous other requirements could be listed, however they deal mainly with the technology in its present and near future development.

We must be careful that we do not try to force-fit the present day technology into areas that really require advanced speech processing concepts. For example, although it is true that isolated word recognition can solve a wide range of present day problems, care must be taken to discern the limitations of these systems with respect to a truly correct speech application.

The present day focus of attention is in the area of applications rather than basic speech research. The speech pendulum again reached an extreme position, limiting the amount of speech research.

We reiterate that more basic research into speech perception, linguistics, acoustic-phonetics is necessary before a thoroughly adaptive error free speech processing system can be constructed.

BIOGRAPHICAL SKETCH

Dr. Bruno Beek

Dr. Bruno Beek was born in White Plains, NY on July 9, 1931. He received the B.A. degree in physics and mathematics, and the M.S. and Ph.D. degrees in physics from Syracuse University, Syracuse NY, in 1956 and 1970, respectively. In 1956 he joined the Rome Air Development Center, Griffiss Air Force Base, Rome, NY as Research Physicist, doing work in electronic warfare technology. In 1962 he began doing research in automatic speech processing techniques. He has also participated in contract programs related to speech research and has been involved in the Department of Defense long-range planning program. In 1976 he served as a Speech Specialist to the AC/243 (Panel III) Research Study Group 4 in automatic pattern recognition under the North Atlantic Treaty Organization (NATO). He has contributed a number of articles, been invited to participate in national and international conferences and seminars. Dr. Beek is a member of the IEEE S-ASSP Speech Processing Technical Committee.