# An Assessment of the Technology of Automatic Speech Recognition for Military Applications

BRUNO BEEK, MEMBER, IEEE, EDWARD P. NEUBERG, AND DAVID C. HODGE

*Abstract*—The objective of this paper is to provide a summary of the state-of-the-art of automatic speech recognition (ASR) and its relevance to military applications. Until recently, speech recognition had its widest application in the development of vocoders for narrow-band speech communications. Presently, research in ASR has been accelerated for military tasks such as command and control, secure voice systems, surveillance of communication channels, and others. Research in voice control technology and digital narrow-band 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.

In coordination with the above-mentioned emphasis in military-supported research, this paper is divided into two major sections. The first section presents discussion of the state-of-the-art and problems in the various subareas of the ASR field. The second section presents a number of unsolved problems and techniques which need to be perfected before the solutions to a number of military applications of the ASR field are possible.

## INTRODUCTION

THE OBJECTIVE of this paper is to provide a summary of the state-of-the-art of automatic speech recognition (ASR) and its relevance to military applications. Table I lists a number of military tasks for possible automation using speech recognition technology. Until recently, speech recognition has had its widest application in the development of vocoders for narrow-band speech communications. Presently, research in ASR has been accelerated in the following areas: speech understanding systems, voice control systems, automatic speaker verification (ASV), and narrow-band digital secure voice systems.

These areas, with the exception of speech understanding research, offer the potential for either immediate or near-term solution to a number of military problems. Research in voice control technology and digital narrow-band systems is 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.

Today, only a small number of applications exist for using ASR. However, as these systems prove to be reliable and cost-effective, their usage will increase in the near future. Already in use are voice data entry systems for inventory/supply record control [36]. 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. As such, the voice data entry system eliminates manual transcription and keying operations.

TABLE I
MILITARY TASKS FOR POSSIBLE AUTOMATION

1) Security
   - 1.1 Speaker Verification (Authentication)
   - 1.2 Speaker Identification (Recognition)
   - 1.3 Determining emotional state of speaker (e.g., stress effects)
   - 1.4 Recognition of spoken codes
   - 1.5 Secure access voice identification, whether or not in combination with fingerprints, facial information, identity card, signature, etc.
   - 1.6 Surveillance of communication channels
2) Command and Control
   - 2.1 System control (ships, aircraft, fire control, situation displays, etc.)
   - 2.2 Voice-operated computer input/output (each telephone a terminal)
   - 2.3 Data handling and record control
   - 2.4 Material handling (mail, baggage, publications, industrial applications)
   - 2.5 Remote control (dangerous material)
   - 2.6 Administrative record control
3) Data Transmission and Communication
   - 3.1 Speech synthesis
   - 3.2 Vocoder systems
   - 3.3 Bandwidth reduction or, more general, bit-rate reduction
   - 3.4 Ciphering/coding/scrambling
4) Processing Distorted Speech
   - 4.1 Diver speech
   - 4.2 Astronaut communication
   - 4.3 Underwater telephone
   - 4.4 Oxygen mask speech
   - 4.5 High "*G*" force speech

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

*1) Digital Narrow-band Communication Systems:* A massive effort is underway to develop and implement an all-digital communication system [59].

*2) Automatic Speaker Verification:* An advanced development model has been fabricated for secure access control applications [14].

*3) Training Systems:* A limited speech understanding system is under study for use as a component in a military training system [20].

*4) Distorted Speech Processing (Helium Speech):* Helium speech unscramblers are being developed to allow for adequate diver-to-diver or diver-to-surface communications.

*5) On-Line Cartographic Processing System:* Studies are

Manuscript received April 20, 1976; revised March 21, 1977.

B. Beek is with the Rome Air Development Center, Griffiss Air Force Base, Rome, NY.

E. P. Neuberg is with the Department of Defense, Washington, DC.

D. C. Hodge is with the U.S. Army Human Engineering Laboratory, Aberdeen Proving Ground, MD.

underway to use speech recognition and voice response techniques with a cartographic point and trace processing systems [5]. Rome Air Development Center implemented an operational system for the Defense Mapping Agency (DMA) where bathymetric depth measurements are digitized using a real-time speaker-dependent, isolated word recognition system.

*6) 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 [58].

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

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

This paper is divided into two major sections. The first section presents a discussion of the state-of-the-art and problems in the various subareas according to the outline of applications given in Table I.

The second section (summary) presents a number of unsolved problems and techniques required to be perfected before solutions to a number of military applications are possible.

## I. Military Applications of ASR: State-of-the-Art

### A. Security Applications

Automatic speech recognition may soon be extensively used in the area of security. Present and future military requirements shall place greater emphasis in this area as some of the critical technology is solved. First and foremost in terms of military applications is the problem of automatic speaker verification/identification [7], [22]. This security problem and others are in different stages of solution, some of which are discussed below.

*1) ASV (Authentication):* Speaker verification means the verification or rejection of an individual based on his speech patterns [11], [14], [53]. 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 by the system, 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 them with its reference file, and makes a decision either to accept or reject the individual or requests additional data. ASV technology has a number of advantages that are not always available in other speech recognition problems. Namely, the speaker is cooperative; he is attempting to gain access to some function and hence will be on his best behavior. The speech data spoken by the individual are 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 (S/N) 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 (i.e., not loud enough, garbled, etc.), request individuals to repeat. Furthermore, the communication channel can easily be made identical for both reference and test.

A number of techniques are being investigated by Bell Laboratories, North American Rockwell, and others. Perhaps the most successful is that of Texas Instruments (supported by the United States Air Force/RADC) [14]. Their technique is as follows: each speaker, or potential user, prerecords a set of 16 words from which 32 sentences can be generated. Each word was selected to contain either a different vowel or diphthong. Only the vowel or diphthong region of the word is used for the speaker verification task. In each of these sentences, critical regions are chosen at which the overall amplitude is large (so that measurements are reliable) and some phonetic event seems to be occurring. These places are found by taking a Fourier spectrum every 10 ms, and looking for regions where spectra are changing rapidly with time (i.e., a transition). For each such transition, a matrix is formed which contains energies in some number of frequency bands from 300 to 2500 Hz, and at times for 100 ms surrounding the event. To score a reference against a test sentence, the incoming sentence is first Fourier transformed every 10 ms. Each template is scanned through the test sentence at 10 ms intervals and at each position is compared with the corresponding 100-ms-long matrix at that position; the distance measure is simply the sum of the squares of the differences between template elements and corresponding elements from the six minima is the score (against those six templates) for the sentence. If the score is below some threshold, the sentence is accepted (and thus the speaker is accepted).

This technique now produces results on the order of 1 percent rejection of true speaker, 2 percent acceptance of impostors, based on only one reference utterance for 120 speakers. Using two reference utterances reduces the false acceptance rate to less than 1 percent. Hence, ASV technology has proven highly successful for fixed context speaker verification when using cooperative speakers in a good S/N environment.

In addition, the technology has successfully coped with the problems 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 acceptances) 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 (i.e., utter two or three phrases instead of one).

The technology has been tested using a large data base and achieves low error rates with a high confidence level. The problem now is to optimize this technology so that it is compatible and acceptable to the user. The user must have complete confidence in the technology, and the problems involving human factors must be eliminated [21].

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 in the United States for a device to handle this problem under field conditions.

*2) Speaker Identification (Recognition):* This section describes several techniques for automatic speaker identification of an individual based on his voice characteristics [1], [7], [22], [32], [63]. Only those techniques will be discussed that have been evaluated in a set of experiments that encompasses a wide range of conditions. These techniques render recognition decisions automatically from continuous speech uncontrolled as to context [29], [67]. Hence, the techniques have the advantage of generating a speaker's reference library and testing unknown speech samples 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 in monitoring communication channels. Conditions that make this problem difficult in practice or real application are:

1) the communication system is of poor quality;
2) speakers are noncooperative;
3) recording and/or channel conditions are different for references 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.

The techniques for the most part deal with analysis and classification of the speech signal in the frequency domain. Essentially, two different approaches have been extensively investigated:

1) extraction of voiced characteristics from long-term and short-term spectral characteristics [1], [22], [29], [67];

2) recognition of specific types of sounds, such as vowels, nasals, and/or fricatives as a prelude to the extraction of spectral characteristics.

Experiments have shown that long-term statistics of speech which include spectral and pitch distributions are pretty well determined by about 30 s of voiced speech [7], [32]. This technique, under laboratory conditions, has fairly low error rates ($\leqslant$5 percent) for up to 40 speakers. These statistics have not, so far, been upheld in an environment which includes channel noise and distortion.

At the present time, greater 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 [64]. Peaks in the vocal-tract transfer function relate to the natural frequencies of the vocal tract and are called formants. It is postulated that measurements of these formants and other formant-related information may be speaker-specific. 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 a linear least-square, inverse filter formulation to estimate the formant trajectories of selected phonetic events [30], [34], [37], [56]. During this research it has been shown, both statistically and experimentally, which analysis conditions give the most stable or consistent parameters and how those same conditions give the best identification performance.

Much has been learned from past research in the field, but there still remain shortcomings which must be resolved in order to remove constraints on text, number of speakers, etc. There are continuing problems with speaker variability, context, and coarticulation effects (see Table II). Research is continuing in these areas, but only at a low level of effort.

*3) Semi-Automatic Speaker Identification System:* Perhaps the most controversial method of speaker identification is by the visual comparisons of speech spectrograms (also called "Voiceprints"). This method is not automatic, but subjective, and its reliability is disputed by many speech researchers [62].

A speech spectrogram is a three-dimensional plot representing time, frequency, and intensity of a sample of speech. Speaker recognition by visual comparison of spectrograms consists of subjectively matching loosely defined points of similarity (in spectrograms of identical utterances) found from the same person that are not found in pairs of spectrograms from different persons. This criterion prevents replication of the experimental results by independent investigations. As such, the "Voiceprints" have been the subject of controversy in the speech community. The advocates of "Voiceprint" are convinced that sound spectrograms are reliable enough to be admissible in a court of law. The opponent position is that the technique is good in a probabilistic sense, but not good enough to be used in convicting or freeing defendents. The situation at the moment is that the Voiceprint technique, administered only by this self-selected group of experts, is in fact being used in courts of law; however, several speech scientists are making strenuous efforts to bring the shortcomings of the technique to the attention of law enforcement officials.

The United States National Institute of Law Enforcement and Criminal Justice of the Law Enforcement Assistance Administration (LEAA) has undertaken the task of developing a computerized system to replace the subjective expert decision of "Voiceprints" by objective probabilistic decisions based on experiments performed on large speaker populations. A computerized hardware/software system has been developed to perform semi-automatic speaker identification [8]. This system will provide interactive displays to allow an operator to analyze quasi-steady-state portions of particular sounds of interest (phonemes). Analytical investigations have involved processing samples of speech of over 250 speakers from which 35 000 phonetic events were extracted. Algorithms were also developed to compute similarity measures based on various combinations of detected phonemes. It is the goal of this LEAA effort that when the semi-automatic speaker identification system is fully implemented and tested, it will overcome many of the current objections to the utilization of voice identification for courtroom evidence.

*4) Surveillance of Communications Channels:* This topic can include a number of items common to all major tasks listed in Table I. However, only two topics will be discussed due to military interest in these areas.

TABLE II
TECHNIQUES REQUIRING PERFECTION TO AUTOMATE MILITARY TASKS

1) Signal Conditioning

Some processing of speech signals may be necessary to compensate for different characteristics of input channels, such as overall signal level and differential delay. Also, it may be possible to preprocess to improve speech quality, or S/N ratio, or to remove long silences.

2) Digital Signal Transformation

The digitized speech signal is transformed in preparation for the extraction of parameters. Processes used include Fourier and Walsh transforms, correlation, LPC, and digital filtering.

3) Analog Signal Transformation and Feature Extraction

The signal can be transformed by hardware, such as filter banks and correlation devices. Transforms can be digitized for further processing, or parameters and features can be extracted in a continuous manner for presentation to decision networks or algorithms.

4) Digital Parameter and Feature Extraction

Calculations are done on the transformed signal to extract relevant parameters, such as formant tracking, pitch extraction and principle components analysis.

5a) Resynthesis

Speech parameters extracted, as mentioned above, in speech compression systems or stored in voice playback systems, must be transformed into acceptable acoustic speech signals.

5b) Orthographic Synthesis

In the translation of written materials to speech, a number of techniques must be developed. Some of these techniques are similar to those cited in the paragraphs above. One of the most important is the development of speech morphology.

6) Speaker Normalization, Speaker Adaptation, Situation Adaptation

The effectiveness of parameters in carrying relevant speech information depends on characteristics of individual speakers and on operational situations. This could mean that systems must be trained or must adapt to optimize parameters.

7) Time Normalization

In recognition of isolated utterances, normalization is imposed to compensate for local and global differences in speech rate. Both linear and nonlinear schemes can be used.

8) Segmentation and Labeling

Segment boundaries are set, e.g., at points of rapid change, formant positions, voicing, spectral shape or other parameters. Segments may be labeled probabalistically to acoustic-phonetic classes. Prestored knowledge of features and parameters for the various classes of segments are used in the decision.

9a) Language Statistics

Language statistics and partial recognition are used to predict and evaluate words at specific points in an utterance.

9b) Syntax

The grammar of the task is used to predict and evaluate word categories at specific points in an utterance.

9c) Semantics

Knowledge of the task domain is used to predict and evaluate subject matter at specific points in an utterance.

9d) Speaker and Situation Pragmatics

In determining the semantics of speech, certain aspects of the utterances are related to an underlying assumption about what the speaker would generally consider an appropriate response. The development of this type of knowledge is required for speech understanding systems. Knowledge of the situation that gave rise to the speaker's utterances is also required for reliable interpretation and execution of the task to be performed in response to the utterance.

10) Lexical Matching

Strings of linguistic-phonetic elements hypothesized by the linguistic part of the system are compared with strings of acoustic-phonetic elements derived from an utterance. A quantitative goodness of match is calculated.

11) Speech Understanding

All sources of knowledge (acoustic, phonetic, pragmatic, semantic, syntactic) are used in combination to reconstruct the utterance and/or determine its meaning.

12) Speaker Recognition

Speaker-specific parameters are extracted and compared with stored parameter sets from known speakers.

13) System Organization and Realization

Systems must be developed keeping in mind use by humans and cost-effective factors.

14) Performance Evaluation

Present development of all speech systems requires the determination of the quantitative value of each possible technique studied. Only by the use of stored speech samples is this performance evaluation possible.

*a) Keyword Classification:* The goal of this program 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, 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.

Two techniques have been investigated. The first 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. Another approach is to simply measure energy in selected frequency bands and match them against stored templates of the same measurement made on the keyword. Neither approach has yet proven successful, especially if the system has not been trained by the speaker. Considerable cost and effort have already been expended on research in this area; results have not been very promising. An experiment was performed using speech from a news-broadcast radio station as input; here the speech is carefully enunciated, the speakers are few, the communication channel is good, and there are frequently occurring words (such as the name of the radio station) [61]. In the best case, there was 80–90 percent recognition of the keyword, with nonnegligible false alarms. After a hiatus, military-sponsored research in this field is recommencing.

*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. Most prominent features have been extracted from long-term speech spectra (usually voiced speech). The features are then used to generate statistics for the various languages of concern. Experimental tests using these statistics as representing the languages have proven that these approaches will not work, or they are too speaker-dependent to be useful. 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 the speech understanding group in 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 linguistic chains of this type, formed from a phonetician's hand transcription of speech in various languages [40]. 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 1–2 min of speech.

Experiments were also performed on five languages involving over 100 different speakers using acoustic events extracted from continuous speech [28]. The data on each speaker were collected over a period of time to determine the effects of speaker variability on recognition performance. Fairly encouraging language recognition scores for three of the languages have already been attained. However, the other two languages were quite poorly recognized, indicating the need for further work. Experiments were also conducted on detecting a single language from many different languages. In one case, a language was detected successfully against several languages involving over 100 speakers with practically perfect detection and no false alarms [28].

### B. Command and Control

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

1) limited word sets
2) connected word recognition (limited set of words)
3) continuous speech recognition/understanding.

*1) Isolated Word Recognition:* Speaker-dependent, limited vocabulary, isolated word recognition (10-50 words) has, for all intents and purposes, been solved under laboratory conditions [4], [6], [9], [25], [31], [33], [35], [42], [44], [46]-[49], [66]. A number of commercial companies in the United States and England are marketing isolated word recognizers. Voice data entry systems are already operating in a variety of industrial applications [36]. These include:

1) automated sorting systems, distribution of parcels, containers, and baggage
2) voice programming for machine tools
3) inspection system, i.e., measurement of TV face plates.

In general, automatic recognition processes can be viewed as a series of algorithms which reduce the relatively high-dimensional measurement space (time domain) to a single-dimensional recognition space. Each transformation, in turn, reduces the number of features that characterize the event. Ultimately, a minimum set of features is extracted and is used to generate a recognition pattern. The final transformation associates the pattern with one of a finite number of patterns of known events.

The most important problem to be solved in the isolated word situation, it seems, is locating the beginning and end of the word [36], [51]. According to many researchers in the field, if the word endpoints can be found, precisely, the exact process used in recognition becomes less critical. The beginning is usually easier to find on the basis of total energy in the signal. However, the end is often masked by breath noise after the word is over, and sophisticated speech/nonspeech separation techniques are needed. Once the endpoints are found, one extracts for the duration of the word, either continuously or at discrete times, some functions of the acoustic signal that are linguistically significant. In one system, for example, 32 functions are calculated continuously (by hardware), varying from something as simple as energy spectrum regions to a function indicating presence of a particular vowel (calculated by detecting voicing, and then looking at the energies in the frequency bands where the formants of that vowel should be found) [35]. These 32 features are sampled at 16 equally-spaced places throughout the word, and the resulting $16 \times 32$ matrix is used as a template to represent the word. In most systems, several tokens of each word are used to "train" the system; some use an average template to represent the word; others use templates for each word [35]. In the recognition mode, the matrix for an incoming word is calculated and matched against all the stored templates; the system decides that the incoming word is that word whose template is closest, according to some metric, to the new matrix. In most systems there is a distance threshold on the distance functions, and if the template fails to meet the distance criteria the word is rejected.

The difficult problem in isolated word recognition seems to be proper alignment of endpoints. Nonlinear time normalization does not seem to be necessary (especially if several training templates are stored), and although the particular functions used must obviously have linguistic significance, various systems using different sets of functions give satisfactory results. Some of the limitations of present systems are:

1) vocabulary is small—about 75 words
2) there is no speaker normalization procedure
3) there is no extrapolation of what the system learns on one training word [2) and 3) mean that every speaker must train the system on every word in the vocabulary]
4) the words must be spoken in a discrete manner.

*2) Multispeaker, Isolated Words:* Great emphasis is being placed today on the development of a speaker-dependent isolated word recognition system which operates over a good quality telephone network (bandwidth 300-3500 Hz) [50], [55], [58]. Experiments have shown that a relationship exists between word size, type of vocabulary, and speaker dependence as a function of recognition accuracy [52], [66]. Although most multispeaker recognition systems tend to be insensitive to variations in the rate of speech, problems still exist due to variations in a talker's speech characteristics

and talker-dependent traits. Recent experiments have shown a multispeaker recognition system to be highly accurate in recognizing a small set of words (10-20) even in the context of connected speech [54], [68]. However, in these experiments, adjustments had to be made for an individual's characteristics by using a short learning phase. Also under experimental investigation is a speaker-independent connected digit recognition system to identify four digit code groups. These codes would be used in conjunction with an ASV system in lieu of a keyboard or badge reader. Being restricted to a finite length set of digits, error correcting codes, a few predetermined digits could be used to adapt the system [14].

*3) Short Phrase Recognition:* Item 2) above implies that isolated word recognition techniques are not sufficient for the short-phrase recognition task. There is, in fact, some military-sponsored research in the recognition of short strings of spoken digits and words. The total vocabulary of strings is too large to be handled by the isolated word algorithms, so two new problems must be faced. First, in place of the endpoint problem we have the segmentation problem: word boundaries must be found in connected text. Second, after words are located, the problem exists that words are altered according to context and according to position in the string. (For example, on the last word of a string the speaker usually allows both the amplitude and pitch to fall greatly.) The word-alteration problem might be solved by storage of more templates for each word, or by application of alteration rules; segmentation techniques are not as successful now as isolated word finders at determining word boundaries, even for a very small vocabulary such as the ten digits.

*4) Continuous Speech Recognition (Speech Understanding):* Speech understanding is interpreted to mean the capability to recognize any sentence that can be produced from a known vocabulary (lexicon), with known subject matter (semantics), using known grammatical rules (syntax) [13], [39], [41], [52], [68]. The conventional approach many are trying in order to get to this problem is to restrict lexicon, syntax, and semantics *drastically*, use high-quality speech (i.e., not limited to telephone bandwidth) and allow only one or a small number of speakers. A very restricted domain, for example, that is being investigated by one group is playing chess by voice [15]. (There are several computer programs that will play the user quite a good game of chess.) The lexicon consists of approximately 50 words. Syntax is very simple and semantics is trivial (few moves are even legal). A more complicated and more useful task domain is querying a data base by voice. An example of this is a query system pertaining to the Navy, of which one can ask questions such as "How many ships have length greater than 700 feet?," or "List all cruisers of the United States Navy."

All continuous speech recognition (CSR) systems so far constructed are large computer programs (100 000-200 000 instructions) on fairly large computers, typically a DEC PDP-10. The digitizing is done at 20 000 samples/s in most systems, allowing a bandwidth of about 10 kHz. Pitch extraction is performed in a variety of ways. With very few exceptions, spectral analysis is accomplished via linear predictive coding (LPC) and a Fourier transform, with formant tracking achieved by various ad hoc procedures [2], [3], [27], [50], [52], [68]. Vowel and fricative identification is implemented by matching LPC spectra taken every 10 ms against a stored template, or extracting formants and matching formant positions against formant templates. Stop identification is accomplished by recognizing short silences following bursts, and by observing formant transitions in adjacent vowels. Nasals, liquids, and glides are found by observing voicing with low overall energy, and then matching templates as if for vowels. In some systems, templates are formed by training the system on some specially constructed test sentences, producing a new set of templates for each speaker. In others, an attempt is made to normalize the speaker's parameters so that a single set of templates can be used for all speakers.

At the so-called "top end," the part that hypothesizes words to be matched against the incoming acoustic stream, fairly sophisticated new techniques are being used to decide what words are likely to appear at a given place in the sentence given a tentative partial recognition. Words are generally stored as a string of syllables or phonemes. Before that string is compared with the string of linguistic elements derived from the acoustic stream, it must be altered by application of phonological rules to take into account such things as context, amplitude, and duration changes induced by syntax and semantics, and local variations in the observed partial recognition. Matching algorithms for best recognition are still rather crude, and statistics of phones or phonemes are not used in the decision process. At a guess, the strings of linguistic elements that are being presented to the matching algorithms are something like 40 percent correct when compared to a phonetician's transcription, and probably the correct element is among the first five choices some 80 percent of the time. It is estimated by some that good recognition will require something like 90 percent first choices for the derived linguistic elements, and state-of-the-art is certainly two to three years from that goal [15].

There has never been a stated military requirement for CSR, but as with other processes, it would clearly be beneficial. The Defense Advanced Research Projects Agency (ARPA) has recently supported a very large research effort in CSR. Outside of the Department of Defense (DOD), the largest single effort is at IBM, which has a 15-man effort in progress [13]. The ARPA project culminated in late 1976. Limited progress has been achieved toward speech understanding goals, with the greatest success coming from the HARPY system, developed by Carnegie-Mellon University. The main targets of high accuracy on connected speech with a vocabulary of 1000 words were met by the HARPY system, with accuracy given in terms of total sentence accuracy, 91 percent, and semantic accuracy, 97 percent. Processing by the HARPY system takes about 20 times real time, and although HARPY is not a general task-free system, some constraints were not necessary such as the use of very high quality nonnoisy microphones; also, although HARPY makes substantial use of training, requiring the experimenter to adapt its word dictionaries, the amount of training required to add a new or additional speaker is not excessive [38].

### C. Data Transmission and Communication

Extensive work in the area of speech transmission has been carried out. Generally, the programs tend to follow similar lines of attack, particularly with regard to the new innovation, LPC vocoders. Conventional vocoder systems have been widely used in military and government applications and are still under investigation. The following discussion outlines the United States programs for an all-digital communications system.

Because the United States DOD has decided to revamp its communications and go to an all-digital voice system, while keeping bandwidths low, massive compression techniques are being studied, developed, and realized in hardware and software [59]. Six defense agencies (see Appendix E) are carrying on and/or sponsoring research in the millions-of-dollars-per-year range. The following is a description of the techniques being pursued, and where in the study/research/development/implementation cycle they now stand.

*1) Conventional Vocoders:* Despite the tremendous popularity of the new LPC techniques, conventional vocoders are still being considered as candidates for the system [16]. One of the tests in the grand competition will be comparison with HY-2, which is a (by now) fairly old channel vocoder that has been used a great deal by the military. A new, all-digital, very reliable version of the HY-2 will be built and used in the tests. Besides this, new techniques are being examined to build better channel vocoders. (No other types of vocoders, such as Laguerre or harmonic vocoders, are being studied.) Some improvement seems to be possible in the coding of channel amplitudes; because of high correlation among amplitudes of the higher channels, coding of amplitude relative to the next lower channel is more economical than coding of amplitude itself. There is one development effort that is simply trying to build a classical channel vocoder, paying great attention to detail in order to improve quality. The implementation will in fact be digital, but all processes will be simulations of realizable analog processes, such as Lerner filters and the Gold pitch extractor.

Pitch extraction, which is of course a problem in all vocoders, not only channel vocoders, has improved since the last generation of vocoders, but is still very much an ad hoc process. Much of the activity in pitch detection seems to be an attempt to carry out known pitch detection schemes faster, and on computers with short words (16 bits). One new idea in pitch extraction is use of dynamic programming to track current pitch, based on estimates of past and future pitch.

*2) LPC Vocoders:* The main thrust of the consortium is toward a vocoder using LPC. The idea at its simplest, is to extract parameters through LPC and extract pitch, and reconstruct the input from these data at the other end. The LPC techniques that are being tried are those described by Atal, Markel, and Itakura [3], [27], [34]; these have all been set up in software, and are being realized in hardware. (Hardware here means very fast microprogrammable computers, hard-wired to do the vocoder process. There are a number of these computers on the market, all with some degree of parallel processing and operating in the range of 100 ns/instruction.) Various LPC parameters have been analyzed, such as LPC coefficients themselves, reflection coefficients, and vocal-tract area functions; at present, reflection coefficients are in greatest favor, since they are fairly insensitive to noise. Many different rates of sending parameters have been tried; a good compromise between bit rate and quality seems to be to use 10- or 12- coefficient LPC and to send (update) the parameters every 16 or 20 ms. For smoothing purposes, the parameters are calculated more often and transmitted at the indicated rate.

Some LPC-related ideas that have been considered in the past but not tried are being experimented with, but have not been implemented into hardware or major software. One of these is pitch-synchronous LPC calculation. There seems to be a stability gain in doing this; however, since parameters are being transmitted at regular times, there is some uncertainty as to the best means of encoding. Another technique being examined is Kalman filtering. This is presently in the study stage, but certain problems have already emerged; it is more sensitive to noise than LPC and less stable. A possible use of Kalman filtering is to do short LPC's (on the order of 3 ms) and Kalman filter the coefficients. Both versions of Kalman filtering will have to be done pitch-synchronously.

*3) Other Vocoders:* Some novel hybrid vocoders are being considered, somewhat akin to voice-excited vocoders (VEV) [16], [45]. One such device will be a standard channel vocoder from 1-3 kHz, but the first 1000 Hz will be encoded by a fourth-order LPC calculation. Another will be a fairly standard VEV, in which the bottom 800 Hz are encoded by adaptive pulse-code modulation (ADPCM) or adaptive delta modulation (ADM). Still a third idea is to do the LPC calculation on the whole signal to obtain coefficients, then use the residual to convey the pitch; this will be done by filtering the residual by a low-pass filter and either sending it as is or compressing it via ADPCM or delta modulation, or even via a low-coefficient LPC.

The ARPA part of the consortium effort has some rather special aspects to it, since ARPA is concerned not only with speech compression, but also with the use of networks in communication. They have a network in which information is sent in 1000-bit packets, each containing the address of the recipient. These are switched at connection points in the network to whatever path toward the recipient is open at the time of arrival of the packet. Since different paths may cause different delays, the packets must be stored and reassembled at the receiving end. Clearly, this causes problems in real-time applications such as a voice communication system, and much of the ARPA work addresses these problems. Besides these special network problems, the ARPA effort is looking at another aspect of compression, which is removal of redundancy after a process like LPC or channel vocoding. There is some hope that by very careful editing of the redundant portion of the parameterized speech (especially silence), and accepting the resulting transmission delay, one can get high-quality vocoded speech at 1000 bits/s.

It is interesting to note that, except for this last notion of removal of redundancy, all compression efforts have a common underlying principle; they all try to separate the speech into a source function and a transfer function, encode these

two separately, and reconstruct at the other end by convolution. No other model of speech compression is being considered.

### D. Processing Distorted and Degraded Speech

*1) Distorted Speech Process (Helium Speech):* The use of a helium-rich mixture as a breathing gas solves physiological problems in the deep-sea or saturation diving situation, but gives rise to a problem of helium speech, the degradation of speech intelligibility [18], [54]. There is no question that effective speech communication must be provided for life support and for the effectiveness of divers in such situations. However, the present state-of-the-art for overcoming the problem is unsatisfactory.

Many attempts have been made to clarify the nature of helium distortions and then to build helium-speech unscramblers. Except for a few of them, almost all seem to have some shortcomings in understanding the underlying nature of helium distortions. It is due to: 1) the lack of insight into the basic mechanism of speech production and perception, 2) the focusing of attention only on narrow questions and on the small sets of data, and 3) the nonutilization of common knowledge and well-known techniques in the field of speech processing.

"Helium speech" occurs when divers perform saturation dives to depths of 200 ft and greater. An oxygen-helium mixture is breathed at pressures up to 450 lb/in$^2$. Marked changes in the resonant characteristics of the vocal tract result. The generalized effects of these conditions on speech are as follows.

1) There is a linear transformation of formant frequency above 500 Hz.

2) There is a nonlinear transformation of formant frequency below 500 Hz.

3) There is an apparent loss of energy in consonants compared to vowels.

4) Formant bandwidths do not increase with increasing pressure.

5) High-frequency energy decreases with increased pressure.

6) Low-frequency energy increases with increased pressure.

There has been little investigation of the effects of helium and pressure on vowel-consonant transitions. Behavioral characteristics of divers add to the problem. Some, for example, try to speak louder to overcome high ambient noise and this affects the consonant-vowel ratio further. The configuration of the mask and other life-support equipment also contributes to the problem.

The general approach taken in trying to unscramble helium speech involves linear shifting the vowel format frequencies down to the "normal" positions. This approach ignores the nonlinear low-frequency problem. The methods used fall into two categories: frequency domain processing and time domain processing. Frequency domain processing includes subtraction and vocoding. Time domain processing includes tape recorder manipulation, convolution processing, and digital coding schemes.

Typical examples of the two main methods used in unscrambling helium speech are given as follows:

*a) Frequency Domain Processing*

*i) Frequency Subtraction:* This involves subtracting a fixed frequency from the entire helium-speech spectrum [10]. It is generally accomplished by heterodyning a bandpassed version of the incoming signal by a carrier frequency, selecting one sideband, and then heterodyning it down in frequency. Another approach is to split the incoming signal into two subbands using bandpass filters. These bands are heterodyned up by a carrier frequency, then heterodyned downward by separately tuned oscillators. Frequency subtraction has been used in unscramblers developed by the Naval Applied Sciences Laboratory.

*ii) Vocoder Techniques:* Formant-restoring vocoders are used to compress the speech spectral envelope while preserving the harmonic structure of the speech signal [19]. Helium speech is introduced into a contiguous bank of bandpass filters whose outputs are full-wave rectified and smoothed by low-pass filters, yielding a slow time-varying signal that is proportional to the energy in the passband of a given analyzing filter. These signals are used to balance modulate an excitation signal derived from the original helium speech, and then are resynthesized through a bank of bandpass filters. This technique has been implemented in an off-line mode only.

*b) Time Domain Processing*

*i) Tape Recorder Playback:* This was one of the first techniques attempted to unscramble helium speech, and it involves recording the helium speech at one speed and then playing it back at one-half speed [24]. Off-line processing is required to restore the original time base. More recently, sophisticated recording and playback techniques, including revolving tape heads, etc., have been attempted.

*ii) Digital Coding:* Digital circuitry has been used in unscramblers built by Raytheon Company, Industrial Research Products, Inc., Westinghouse, and Singer. In general, the helium speech is sampled in real time, stored in a register, and then read out at a slower rate while bandpassing the signal. Some information is discarded in order to achieve proper frequency scaling. (Further details of these devices are proprietary, and hence not available.)

*iii) Convolution Processing:* This approach takes advantage of the fact that the shape of the vocal tract is the same under both helium and air speaking. The helium-speech waveform is first deconvoluted to obtain vocal-tract impulse response functions. Then an inverse can be constructed, and appropriate time scaling can be performed to compensate for the necessary changes in the velocity of sound.

On several of the existing scramblers, tests have been performed using rigorous and systematic procedures under a United States Navy contract by Hollien *et al.* [54] at the University of Florida. The results of these indicated that, in general, none of the unscramblers evaluated to date provides enough substantial improvement to allow for really adequate diver-to-diver or diver-to-surface communication. Mean unprocessed PB word intelligibility was reported to be about 16 percent, and mean unscrambled PB intelligibility was no better than about 40 percent. (Typical current military communication system criteria require 70 percent PB

word intelligibility as mandatory and 80 percent as desirable.) Single digits can, of course, be received reliably with systems having PB intelligibility on the order of 50 percent, but good sentence intelligibility requires much higher values.

In general, it seems that helium-speech unscramblers do not work very well (although they may be in use because nothing better is available). Unsolved problems include the nonlinear shift at lower frequencies, formant transitions and consonants, and restoring consonant-vowel energy ratios. Hybrid systems involving both time and frequency domain processing are probably the next step in the development process.

In conclusion, a comprehensive study covering a wide range of speech phenomena is indispensable as it will lead to the following advantages: 1) providing useful data for designing a reasonable and practical unscrambler, 2) providing insight into the basic mechanism of speech production in media different from normal air, and 3) providing insight into the perceptual mechanism of speech in such media. Such studies are necessary for the improvement of helium speech.

*2) Processing Degraded Speech:* 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 nonspeech 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 wide-band and random. It is similar to homomorphic filtering, with the spectrum rooted (rather than logged) before being transformed. Test results showed the S/N ratio was somewhat improved without seriously distorting the character of the speech [65].

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 attenuating 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 intelligible [65].

The enhancement process 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.

A theoretical framework has been developed for the enhancement system, and meaningful experiments are underway [43]. While the enhancement process has been designed principally to enhance a voice signal immersed in wide-band noise, it is also providing insight into the enhancement of a voice signal subject to narrow-band noise.

*3) Processing Stressed Speech:* Other voice changes occur as a result of working in particular environments or being subjected to stressful (e.g., emergency) situations. For example, an aircraft pilot's voice characteristics may change as a result of *G*-forces. Threat, or other stress, may cause emotional changes in voice characteristics.

The state-of-the-art in this area appears to be that we are still investigating the extent of these changes in voice quality. Such research is being conducted largely by the DOD; civil projects seem to be largely ignoring these problems.

One project is being conducted by the United States Navy. Investigators are utilizing a dynamic flight simulator to study the effects of adverse flight conditions on voice quality. Typical variables include cockpit noise, high-*G* maneuvers, buffeting, mask breathing, cockpit temperature, and long-duration missions. Word sets being used include those expected to occur in communicating normal and emergency messages. The results will be used to determine: 1) conditions under which existing systems can be utilized, and 2) the extent of adaptability which must be designed into future speech recognition systems.

Stress-induced changes in voice quality may be studied in several ways. One way is to use task-induced stress. The subject is given a heavy workload and is required to make spoken responses. A paper by Hecker *et al.* suggests that knowledge of task-induced changes in voice quality is fairly complete [23].

Emotional stress-induced voice changes are more difficult to determine. An important aspect of this problem is how to induce emotional stress. Some study has been made of radio communications from pilots experiencing trouble. (In at least one case, the pilot crashed and was killed.) However, reference utterances are usually not available, and there is a lot of noise on the tapes. Simulated emergencies have been tried, but no one knows whether the extent of voice changes would be the same in a real emergency. A recent Army program [12] sought to attack this problem in the following way. Paid volunteer subjects were placed under hypnosis and lead via suggestion to relive traumatic experiences from their pasts. In one case, a man who had been a bomber crewman relived his experience of trying to release a live, 500-lb bomb that had become jammed in the bomb-bay. Speech was recorded before, during, and after such sessions. Data from this project are expected to provide evidence of stress-induced speech changes that must be taken into account in designing speech recognition systems.

## II. Summary of Techniques to be Perfected and Unsolved Problems

Since its inception, research in ASR has progressed a long way. Man's most natural means of communication has undergone considerable analysis, resulting in a massive amount of equipment and documentation. Models of the speech mechanism have been developed and tested to demonstrate basic feasibilities. Earlier techniques have been considerably improved. Numerous independent and integrated efforts have been undertaken to further develop the speech genera-

TABLE III
STATE-OF-THE-ART AND UNSOLVED PROBLEMS

| Processing Techniques | State-of-the-Art[1] | Unsolved Problems[2] |
|---|---|---|
| 1) Signal Conditioning | A, except speech enhancement (C) | 1, 15, 20, 23 |
| 2) Digital Signal Transformation | A | 1, 15, 20 |
| 3) Analog Signal Transformation and Feature Extraction | A, except feature extraction (C) | 1, 2, 6, 14–16, 20 24, 25 |
| 4) Digital Parameter and Feature Extraction | B | 1, 2, 6, 14, 16, 20 24, 25 |
| 5a) Resynthesis | A | 4, 7, 20, 25 |
| 5b) Orthographic Synthesis | C | 4, 6–8, 19, 26–28, 29 |
| 6) Speaker Normalization Speaker Adaptation Situation Adaptation | C | 15–17, 19, 20, 23–25, 29 |
| 7) Time Normalization | B | 3, 16, 20, 25, 29 |
| 8) Segmentation and Labeling | B | 1, 4, 5–9, 11, 13, 16, 18–20, 24, 25 |
| 9a) Language Statistics | C | 5, 8, 9, 11, 12, 14, 20, 24, 25 |
| 9b) Syntax | B | 6, 7, 9, 12, 14, 20, 25 |
| 9c) Semantics | C | 6, 7, 9, 12, 14, 20, 25, 10 |
| 9d) Speaker and Situation Pragmatics | C | 3, 6, 12, 14, 16, 18, 19, 23 |
| 10) Lexical Matching | C | 7–9, 12–14, 20, 25 |
| 11) Speech Understanding | B–C | 5, 9, 12, 14, 16, 18, 20, 23, 25 |
| 12) Speaker Recognition | A for speaker verification; C for all others | 14, 16, 17, 19, 20, 24, 25 |
| 13) System Organization and Realization | A–C | 21, 22 |
| 14) Performance Evaluation | C | 1, 6–11, 18–20, 24–28 |

[1]Ratings: A = useful now; B = shows promise; C = a long way to go.
[2]See Glossary (Table IV) for problem definitions; list may not be exhaustive.

TABLE IV
GLOSSARY OF PROBLEMS AND DEFINITIONS

1) Detect speech in noise; speech/nonspeech.
2) Extract relevant acoustic parameters (poles, zeros, formant (transitions), slopes, dimensional representation, zero-crossing distributions).
3) Dynamic programming (nonlinear time normalization).
4) Detect smaller units in continuous speech (word/phoneme boundaries; acoustic segments).
5) Establish anchor point; scan utterance from left to right; start from stressed vowel, etc.
6) Stressed/unstressed.
7) Phonological rules.
8) Missing or extra added ("uh") speech sound.
9) Limited vocabulary and restricted language structure necessary; possibility of adding new words.
10) Semantics of (limited) tasks.
11) Limits of acoustic information only.
12) Recognition algorithm (shortest distance, (pairwise) discriminant, Bayes, probabilities).
13) Hypothesize-and-test, backtrack, feed forward.
14) Effect of nasalization, cold, emotion, loudness, pitch, whispering, distortions due to talker's acoustical environment, distortions by communication systems (telephone, transmitter-receiver, intercom, public address, face masks), nonstandard environments.
15) Adaptive and interactive quick learning.
16) Mimicking; uncooperative speaker(s).
17) Necessity of visual feedback, error control, level for rejections.
18) Consistency of references.
19) Real-time processing.
20) Human engineering problem of incorporating speech understanding system into actual situations.
21) Cost-effectiveness.
22) Detect speech in presence of competing speech.
23) Economical ways to adding new speakers to system.
24) Use of prosodic information.
25) Coarticulation rules.
26) Morphology rules.
27) Syntax rules.
28) Vocal-tract modeling.

tion, reception, and reproduction phenomena for specific military and civilian applications. This has increased the interaction among scientists of various disciplines. These include interchange and interaction in acoustic-phonetics, linguistics, signal processing, computer science, etc. In fact, international participation in the solution of the numerous ASR problems may also be at hand. European laboratories engaged in ASR are concentrating on statistical classification and theoretical models of detection, and are presently less involved in artificial intelligence techniques. Combining the various technologies could also lead to positive international cooperation.

A number of techniques must be perfected (and problems to be solved) before automation of the military tasks shown in Table I can be achieved. These techniques, together with their definitions and the state-of-the-art, are listed in Tables II and III. A glossary of problems and definitions is given in Table IV.

Although a great deal of emphasis is presently placed on the development of ASR systems for military applications, a much greater understanding of the speech process is required before ASR can approach human performance. Significant advances in ASR are not likely to come solely from research in pattern recognition (artificial intelligence) computer development and waveform/signal processing. Although these areas of investigation are important tools for ASR, the significant advances will come from studies in acoustic-phonetics, speech perception, linguistics, and psychoacoustic equipments.

## APPENDIX A

## FRENCH AGENCIES INTERESTED IN ASR

*1) Government Laboratories*

Ministry of Telecommunications
- Centre National d'Etudes des Telecommunications
- Ecole Nationale Superieure des Telecommunications

Ministry of Defense
- Laboratoire Central de l'Armement

Ministry of Education
- Faculte d'Orsay
- Faculte de Toulouse

Faculte de Nancy
Faculte de Grenoble, Institut de Phonetique
Commissariat a l'Energie Atomique

*2) Firms*

Thomson-CSF, Cagnes sur Mer
Thomson-CSF, Genevilliers
CGE-CIT
IBM-France

## Appendix B

### German Agencies Interested in ASR

Institut fur Kommunikationsforschung und Phonetic
Universitat Bonn
5300 Bonn
Adenauerallee 98a

Allgemeine Flektrizitats-Gesellschaft AEG-Telefunken
Bereich Forschung
7900 Ulm
Postfach 830

Lehrstuhl und Institut fur Allgemeine Nachrichtentechnik
Universitat Hannover
3000 Hanover
Callinstrabe 15

Fernmeldetechnisches Zentralmt
6100 Darmstadt
Postfach 800

Institut fur Nachrichtentechnik
Technische Universitat Braunschweig
3300 Branunschweig
Muhlenpfordstabe 23

Fachgebiet Ubertragungstechnik im Fachbereich Nachrichtentechnik
Technische Hochschule Darmstadt
6100 Darmstadt
Merckstrasse 25

Philips Forschungslaboratorium Hamburg Gmbll
Abt. Datentechnik
2000 Hamburg 54
Postfach 540840

Heinrich-Hertz Institut
1000 Berlin

Institut fur Datenverarbeitung
Technische Universitat Munchen
8000 Munchen 2
Arcisstrasse 21

Institut fur Phonetik und Sprachliche Kommunikation
Universitat Munchen
8000 Munchen 13
Schellingstrasse 7/11

## Appendix C

### Dutch Agencies Interested in ASR

Postal Research Laboratory (Dr. Neher-Laboratory)
St Paulusstraat 4,
Leidschendam
Speaker Verification (Drs. H. R. C. Tromp)

Institute for Perception TNO, Kampweg 5, Soesterberg
Analysis, Synthesis, Perception, and Automatic Recognition of Speech (Ir. L.C.W. Pols)

University of Nijmegen, Institute for Phonetics,
Erasmuslaan 40, Nijmegen
Speech Research (Prof. Dr. W. H. Vieregge)

Industrial Laboratory, Philips, Geldrop
Word Recognition (I. H. Slis)

Delft University of Technology, Department of Applied Physics, Subgroup on Pattern Recognition, Lorentzweg 1, Delft
All Aspects of Pattern Recognition (Prof. Dr. C. J. D. M. Verhagen)

Delft University of Technology, Department of Electrical Engineering, Laboratory for Information Theory, Mekelweg, 4, Delft
Speech Analysis and Recognition (Ir. C. Kamminga)

## Appendix D

### United Kingdom Agencies Interested in ASR

University of Keele
North Stafford Polytechnic
University of Essex, Department of Engineering
University of Essex, Language Centre
University College, London
Combridge University
Manchester University

## Appendix E

### United States Agencies Interested in ASR

*1) DOD Research Laboratories*

U.S. Army Electronics Command, Ft. Monmouth, NJ
Aerospace Medical Research Laboratory, Wright-Patterson AFB, Wright-Patterson, OH
AF Avionics Laboratory, Wright-Patterson AFB, Wright-Patterson, OH
Rome Air Development Center, Griffiss AFB, Rome, NY
Naval Electronics Laboratory Center, San Diego, CA
Naval Air Development Center, Warminster, PA
Naval Research Laboratory, Washington, DC
Naval Ship R&D Center, Bethesda, MD
Naval Training Equipment Center, Orlando, FL
Naval Undersea Center, San Diego, CA
Naval Underwater Systems Center, Newport, RI
National Security Agency, Ft. Meade, MD
Defense Communications Agency, Reston, VA

*2) DOD Monitoring Agencies*

U.S. Army Research Office, Durham, NC
AF Office of Scientific Research, Arlington, VA
AF Electronic Systems Division, Bedford, MA
Office of Naval Research, Arlington, VA
Defense Advanced Research Projects Agency, Arlington, VA

*3) Consortium for All-Digital Communications System Research*

National Security Agency, Ft. Meade, MD
Naval Research Laboratory, Washington, DC
Defense Communications Agency, Reston, VA
U.S. Army Electronics Command, Ft. Monmouth, NJ
Air Force Electronic Systems Division, Bedford, MA
Defense Advanced Research Projects Agency, Arlington, VA

*4) Contractors for ARPA-Sponsored Research on Continuous Speech Understanding*

Lincoln Laboratory, Lexington, MA
Bolt Beranek and Newman, Inc., Cambridge, MA
Carnegie-Mellon University, Pittsburg, PA
Stanford Research Institute, Menlo Park, CA
Systems Development Corporation, Santa Monica, CA
Sperry-Univac, Minneapolis, MN

## REFERENCES


[1] B. S. Atal, "Automatic recognition of speaker from their voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460-475, 1976.

[2] —"Linear prediction of speech–recent advances with applications to speech analysis," in *Speech Recognition: Invited Papers of IEEE Symposium*. New York: Academic, 1975.

[3] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.* vol. 50, pp. 637-655, 1971.

[4] D. Becker, "Automatische Worterkennung mit Einem Linearen Quadratmittel-Klassifikator," in *Proc. NTG-GI Fachtagung*, "Cognitive Verhafen und System," Th. Einsele, W. Giloi, and H. Nagel, Eds. Berlin: Springer-Verlag, 1973, pp. 145-159.

[5] B. Beek, "Applications of speech processing for military applications," A. L. Gilbert, Ed., in *Proc. Workshop on Military Applications of Artificial Intelligence*, White Sands Missile Range, Juarez, Mexico, Oct. 1976, pp. 302-318.

[6] B. Beek and R. Vonusa, "An automatic isolated word recognition system–study and evaluation," Rome Air Dev. Cen. Tech. Rep. RADC-TR 70-129, Rome, NY, Oct. 1971.

[7] B. Beek *et al.*, "Automatic speaker recognition system," presented at the Agard Conf. on Artificial Intelligence, no. 94, London, England: Harford House, 1971.

[8] P. K. Broderick, J. E. Paul, and R. J. Renmick, "Semi-automatic speaker identification system," in *Proc. Carnahan Conf. Crime Countermeasures*, Univ. Kentucky, Lexington, 1975.

[9] R. Carre, "A summary of speech research activities in France," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 268-272, Aug. 1974.

[10] M. Copel, "Helium voice unscrambling," *IEEE Trans. Audio Electroacoust.*, vol. AU-14, pp. 122-126, Sept. 1966.

[11] S. K. Das and W. S. Mohn, "A scheme for speech processing in automatic speaker verification," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 32-43, 1971.

[12] J. DeClerk, U.S. Army Electron. Command, Fort Monmouth, NJ, personal communication.

[13] N. R. Dixon and C. C. Tappert, "Toward objective phonetic transcription–an in-line interactive technique for machine-processed speech data," *IEEE Trans. Man-Machine Syst.*, vol. MMS-11, pp. 202-210, 1970.

[14] G. R. Doddington, "Speaker verification, vol. I-III," Rome Air Dev. Cen., Rome, NY, Tech. Reps. TR-74-179, TR-75-274, TR-76-262.

[15] L. D. Erman, Ed., "Contributed papers of IEEE Symposium on Speech Recognition," Carnegie Mellon Univ., Pittsburgh, PA, Apr. 1974, IEEE Catalog No. 74CH0878-9AE.

[16] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd ed. Berlin: Springer-Verlag, 1972.

[17] J. W. Forgie, D. E. Hall, and R. W. Wiesen, "An overview of the Lincoln Laboratory speech recognition system," *J. Acoust. Soc. Amer.*, vol. 56, p. S27 (A), 1974.

[18] T. A. Giordano, H. B. Rothman, and H. Hollien, "Helium speech unscramblers–A critical review of the state-of-the-art," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, no. 5, pp. 436-444, Oct. 1973.

[19] R. M. Golden, "Improving naturalness and intelligibility of helium-oxygen speech, using vocoder techniques," *J. Acoust. Soc. Amer.*, vol. 40, pp. 621-624, 1966.

[20] M. W. Grady and M. B. Herscher, "Advanced speech technology applied to problems of air traffic control," in *NAECON 75 Record 541*.

[21] W. Haberman and A. Fejfar, "Automatic identification of personnel through speaker and signature verification-system description and testings," in *Proc. Carnahan and Int. Crime Countermeasures Conf.*, 1976, pp. 23-30.

[22] M. H. L. Hecker, "Speaker recognition–An interpretive survey of the literature," American Speech and Hearing Association, Mono. 16, Jan. 1971.

[23] M. H. L. Hecker, K. N. Stevens, G. von Bismarck, and C. F. Williams, "Manifestations of task-induced stress in the acoustic speech signal," *J. Acoust. Soc. Amer.*, 1968.

[24] K. Holywell and G. Harvey, "Helium speech," *J. Acoust. Soc. Amer.*, vol. 36, pp. 210-211, 1964.

[25] O. Hinrichs and J. Gonschorek, "Ein Spracherkennungsgerat mit Selbsttatiger Anpassung an Sprechgeschwindigkeit und Lautstarke," *Nachrichtentech. Z.*, vol. 24, pp. 177-182, 1971.

[26] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, Feb. 1975.

[27] F. Itakura and S. Saito, "On the optimum quantization of feature parameters in the Parcor speech synthesizer," Paper L4 in *Proc. 1972 Conf. on Speech Communications and Processing*, Air Force Cambridge Res. Lab., L. G. Hanscom Field, Bedford, MA, 1972, IEEE Cat. No. 72 CHO596-7AE.

[28] R. G. Leonard and G. R. Doddington, "Automatic classification of languages," Rome Air Dev. Cen., Rome, NY, Tech. Rep. TR-75-264, 1975.

[29] K. P. Li and G. W. Hughes, "Talker differences as they appear in correlation matrices of continuous speech spectra," *J. Acoust. Soc. Amer.*, vol. 55, pp. 833-837, 1974.

[30] J. Makhoul, "Linear prediction; A tutorial review," *Proc. IEEE*, pp. 561-580, Apr. 1975.

[31] H. Mangold, "Ein Spracherkennungssystem fur Isolierte Worte unter Schlechten Storbedingungen," presented at the 5th Conf. on Acoust., Budapest, Hungary, Apr. 1973.

[32] —, "The individual differences with long time measurements of the speech signal," *Nachrichtentech. Z.*, vol. 22, no. 6, pp. 364-367, 1969.

[33] —, "Investigation of a system for recognition of isolated words spoken in noisy surroundings," *Nachrichtentech. Z.*, vol. 27, no. 2, pp. 105-108, 1974.

[34] J. D. Markel, "Formant trajectory estimation from a linear least-square inverse filter formulation," Speech Commun. Res. Lab., Santa Barbara, CA, SCRL Mono. No. 7, Oct. 1971.

[35] T. B. Martin, "Applications of limited vocabulary recognition systems," in *Speech Recognition, Invited Papers of the IEEE Symp.* New York: Academic, 1975, pp. 55-71.

[36] —, "Practical applications of voice input to machines," *Proc. IEEE*, vol. 64, pp. 487-501, Apr. 1976.

[37] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 135-141, 1974.

[38] M. F. Medress *et al.*, "Speech understanding: Report of a steering committee," Massachusetts Inst. Technol., Cambridge, to be published.

[39] E. P. Neuberg, "Philosophies of speech recognition," in *Speech Recognition, Invited Papers of the IEEE Symp.* New York: Academic, 1975, pp. 83-95.

[40] E. P. Neuberg, R. E. Wohlford, and A. S. House, "Preliminaries to the automatic recognition of speech," *J. Acoust. Soc. Amer.*, vol. 57, suppl. 1, p. S34, 1975.

[41] A. Newell *et al.*, "Speech understanding systems: Final report of a study group," 1971 (reprinted by North-Holland/American/Elsevier, Amsterdam, The Netherlands, 1973).

[42] D. J. P. J. Nierop, L. C. W. Pols, and R. Plomp, "Frequency analysis of Dutch vowels from 25 female speakers," *Acustica*, vol. 292, pp. 110-118, 1973.

[43] T. W. Parsons and M. Weiss, "Enhancing/intelligibility of speech in noisy or multi-talker environments," Rome Air Dev. Cen., Rome, NY, Rep. TR-75-155, AD-AO13767, 1975.

[44] E. Paulus, "The role of speech sounds in the perception and recognition of words," in *Zeichenerkennung durch Biologische*

*und Technische Systeme*, O. J. Grusser and R. Klinke, Eds. Berlin: Springer-Verlag, 1971, pp. 357-368.
[45] E. Paulus and D. Langle, "Die Anwendung der Karhumen-Loeve-Entwicklung fur die Digital Sprachanalyze und-Synthese," in *Proc. NTG-Fachtagung "Singalverarbeiting,"* W. Schubler, Ed. Berlin: Erlangen, 1973, pp. 362-369.
[46] L. C. W. Pols, "Real-time recognition of spoken words," *IEEE Trans. Comput.*, vol. C-20, pp. 972-978, 1971.
[47] —, "Dimensional representation of speech spectra," presented at the 7th Int. Congr. on Acoust., Budapest, Hungary, 1971.
[48] —, "Segmentation and recognition of monosyllabic words," in *1972 Proc. Conf. on Speech Communication and Processing*, Air Force Cambridge Res. Lab., L. G. Hanscom Field, Bedford, MA, 1972, IEEE Cat No. 72 CHO596-7AE.
[49] L. C. W. Pols, H. R. C. Tromp, and R. Plomp, "Frequency analysis of Dutch vowels from 50 male speakers," *J. Acoust. Soc. Amer.*, vol. 534, pp. 1093-1101, 1973.
[50] *Proc. Journées d'Etude sur la Parole (Annual Study Days on Speech) of Groupement des Acousticiens de Langue Francaise (GALF)*—1st in Grenobl 70, 2nd in Aix en Provence 71, 3rd in Lannion 72, 4th in Brussels 73, 5th in Orsay 74, 6th in Toulouse 75. (*Proceedings* can be obtained from Secretariat du GALF in CNET. Route de Tregastel, 22301 Lannion, France.)
[51] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, no. 2, pp. 297-315, 1975.
[52] D. R. Reddy, "Speech recognition by machine: A review," *Proc. IEEE*, vol. 64, pp. 501-531, Apr. 1976.
[53] A. E. Rosenberg, "Automatic speaker verification systems: A review," *Proc. IEEE*, vol. 64, pp. 475-485, Apr. 1976.
[54] H. B. Rothman and H. Hollien, "Further evaluation of He02 unscramblers under controlled conditions," Rep. CSL/ONR-47 1, July 1972.
[55] M. R. Sambur and L. R. Rabiner, "A speaker independent digit recognition system," *Bell Syst. Tech. J.*, vol. 54, pp. 81-102, 1975.
[56] R. Schafer and E. L. Rabiner, "Digital representation of speech signals," *Proc. IEEE*, vol. 63, pp. 662-677, Apr. 1975.
[57] H. Schunk, "A transceiver from helium atmosphere," *Allg. Elek. Ges. Telefunken*, vol. 61, pp. 378-381, 1971.
[58] M. Simpson, "Word recognition for tactical data systems," A. L. Gilbert, Ed., in *Proc. Workshop on Military Applications of Artificial Intelligence*, White Sands Missile Range, Juarez, Mexico, Oct. 1976, pp. 302-318.
[59] T. Tremain, "Linear predictive coding systems," in *Conf. Rec. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Philadelphia, PA, 1976, pp. 474-478, IEEE Cat. No. 76CH1067-8 ASSP.
[60] R. Turn, A. Hoffman, and T. Lippiatt, "Military applications of speech understanding systems," Defense Advanced Res. Proj. Agency, Arlington, VA, Rep. 1434, AD 787394, June 1974.
[61] V. A. Vitols and J. E. Paul, "An algorithm for detection of keywords in continuous speech," presented at Workshop of Sept. 1971, session III, paper I, Rome Air Dev. Cen., Rome, NY.
[62] "Voiceprint identification," *Georgetown Law J.*, vol. 61, issue 3, Feb. 1973.
[63] H. Wakita, "On the use of linear prediction error energy for speech and speaker recognition," *J. Acoust. Soc. Amer.*, vol. 57, suppl. 1, 1975.
[64] —, "Direct estimation of vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, no. 5, pp. 417-427, Oct. 1973.
[65] M. R. Weiss *et al.*, "Study and development of the intelligence technique for improving speech intelligibility," Rome Air Dev. Cen., Rome, NY, RADC-TR-75-108, 1975.
[66] G. White, "Automatic speech recognition: Linear predictive residual versus bandpass filtering," in *Proc. IEEE Int. Conf. on Cybernetics and Society*, Sept. 1975.
[67] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, pp. 2044-2055, 1972.
[68] W. Woods, "Speech recognition—An overview," A. L. Gilbert, Ed., in *Proc. Workshop on Military Applications of Artificial Intelligence*, White Sands Missile Range, Juarez, Mexico, Oct. 1976, pp. 302-318.
[69] H. Yilmaz *et al.*, "Automatic speaker adaptation," Rome Air Dev. Cen., Rome, NY, Tech. Rep. RADC-TR-76-273, 1976.

# Linear Prediction with a Variable Analysis Frame Size

SATISH CHANDRA, MEMBER, IEEE, AND WEN C. LIN, SENIOR MEMBER, IEEE

*Abstract*—**This paper describes a speech analysis-synthesis system based on stationary linear prediction formulation. This system uses a variable analysis frame size concept. The *k*-parameters are used to represent the spectral information in the speech. The statistical and quantization properties of *k*-parameters are studied in detail. A method for calculating the analysis frame size based on energy and pitch period variations within a speech waveform has been developed. The speech analysis-synthesis system has been implemented on the computing facility of the Signal Processing Laboratory at Case Western Reserve University. Average data rates of 4800, 3600, and 2400 bits/s have been achieved on a limited speech data base of male speakers.**

## I. INTRODUCTION

RECENTLY, linear prediction technique has been used by many speech researchers for the analysis and synthesis of speech waveform. The formant parameters, pitch period, and speech spectrum have been successfully estimated using the linear prediction technique. There are basically two formulations of linear prediction: 1) stationary and 2) nonstationary. Markel's [1] inverse filtering and Itakura and Saito's [2], [3] maximum likelihood estimation methods fall under the category of stationary linear prediction formulation, whereas Atal and Hanauer's [4] linear prediction and Prony's [5], [6] method are categorized as nonstationary linear prediction formulation. The theoretical studies of the relationships between these two formulations have been discussed in some detail by Makhoul and Wolf [7] and by Markel, Gray, and Wakita [8]. The experimental comparison between the two linear prediction formulations, in representing the voiced speech waveform, has been presented in considerable detail by Chandra and Lin [9], [10].

One of the important applications of linear prediction is for efficient and secure speech communication. Many speech analysis and synthesis systems [2], [4], [11]-[13] have been developed based on both linear prediction formulations. All

Manuscript received March 20, 1975; revised March 30, 1976 and December 10, 1976.

S. Chandra is with Bailey Meter Company, a Division of Babcock & Wilcox, Wickliffe, OH 44092.
W. C. Lin is with the Department of Computing and Information Sciences, Case Western Reserve University, Cleveland, OH 44106.