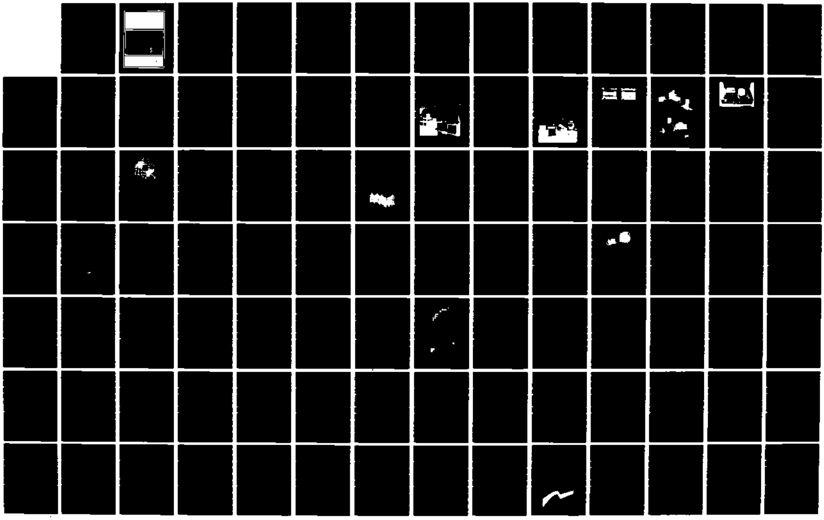
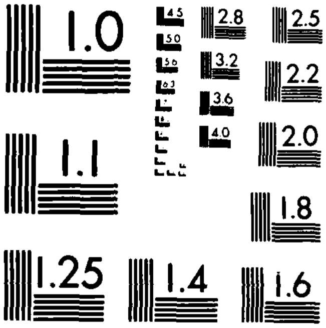
AD-A132 092    SPEECH PROCESSING(U) ADVISORY GROUP FOR AEROSPACE    1/2
RESEARCH AND DEVELOPMENT NEUILLY-SUR-SEINE (FRANCE)
MAY 83 AGARD-LS-129

UNCLASSIFIED    F/G 17/2    NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AGARD-LS-129

AGARD-LS-129

# AGARD

ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT

7 RUE ANCELLE 92200 NEUILLY SUR SEINE FRANCE

AGARD LECTURE SERIES No. 129

# Speech Processing

NORTH ATLANTIC TREATY ORGANIZATION

DISTRIBUTION AND AVAILABILITY
ON BACK COVER

AGARD-LS-129

NORTH ATLANTIC TREATY ORGANIZATION

ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT

(ORGANISATION DU TRAITE DE L'ATLANTIQUE NORD)

AGARD Lecture Series No.129

# SPEECH PROCESSING

The material in this publication was assembled to support a Lecture Series under the sponsorship of the Avionics Panel and the Consultant and Exchange Programme of AGARD presented on 20–21 June 1983 in Trondheim, Norway; on 23–24 June 1983 in Copenhagen, Denmark and on 27–28 June 1983 in Delft, The Netherlands.

## THE MISSION OF AGARD

The mission of AGARD is to bring together the leading personalities of the NATO nations in the fields of science and technology relating to aerospace for the following purposes:

– Exchanging of scientific and technical information;

– Continuously stimulating advances in the aerospace sciences relevant to strengthening the common defence posture;

– Improving the co-operation among member nations in aerospace research and development;

– Providing scientific and technical advice and assistance to the North Atlantic Military Committee in the field of aerospace research and development;

– Rendering scientific and technical assistance, as requested, to other NATO bodies and to member nations in connection with research and development problems in the aerospace field;

– Providing assistance to member nations for the purpose of increasing their scientific and technical potential;

– Recommending effective ways for the member nations to use their research and development capabilities for the common benefit of the NATO community.

The highest authority within AGARD is the National Delegates Board consisting of officially appointed senior representatives from each member nation. The mission of AGARD is carried out through the Panels which are composed of experts appointed by the National Delegates, the Consultant and Exchange Programme and the Aerospace Applications Studies Programme. The results of AGARD work are reported to the member nations and the NATO Authorities through the AGARD series of publications of which this is one.

Participation in AGARD activities is by invitation only and is normally limited to citizens of the NATO nations.

The content of this publication has been reproduced
directly from material supplied by AGARD or the authors.

Published May 1983

ISBN 92-835-0331-7

SPS

*Printed by Specialised Printing Services Limited*
*40 Chigwell Lane, Loughton, Essex IG10 3TZ*

## LIST OF SPEAKERS

Lecture Series Director: Mr J.S.Bridle
Joint Speech Research Unit
Princess Elizabeth Way
Cheltenham, Gloucestershire GL52 5AJ
UK

## SPEAKERS

Dr B.Beek
Rome Air Development Centre
Communications Division
Griffiss Air Force Base
NY 13441
USA

Dr R.Breaux
Naval Training Equipment Center
N-095R
Orlando, Florida 32813
USA

Mr J.R.Costet
Société CROUZET
25, rue Jules Védrines
26027 Valence Cedex
France

Dr M.J.Hunt
National Research Council of
Canada
National Aeronautical Establishment
Building U66
Montreal Road
Ottawa K1A OR6
Canada

Dr H.Mangold
AEG-Telefunken
Research Laboratory
Sedan Strasse 10
D-79 Ulm
West Germany

Dr R.K.Moore
R.S.R.E.
Leigh Sinton Road
Malvern, Worcestershire
UK

## CONTENTS

# GENERAL REVIEW OF MILITARY APPLICATIONS OF VOICE PROCESSING

DR. BRUNO BEEK
MR. RICHARD S. VONUSA

Rome Air Development Center

Griffiss Air Force Base NY 13441

## SUMMARY

This paper introduces the subject of Voice-Interactive Systems and their role in military applications. The history and evolution of automatic speech recognition and synthesis is briefly explored and the current state-of-the-art is reviewed. The term "Voice-Interactive Systems" is defined and the advantages and disadvantages of Voice-Interactive Systems are highlighted. Next, are previous applications of speech systems to military problems is summarized, the major application areas are described and current development projects in the US and other NATO countries are presented. Special attention is focused on the cockpit application. Several projects in this area are discussed along with a summary of important issues to consider when applying Voice-Interactive Systems to the aircraft environment.

## 1. INTRODUCTION

The objective of this paper is to provide a broad perspective on the topic of Voice-Interactive Systems, and in particular the use of these systems for cockpit and other military applications. Before the actual application of these systems is discussed, it would be helpful to consider such questions as: "How did speech systems evolve?", "What techniques are used to do speech recognition and synthesis?", and "What are the major problems in doing recognition and synthesis?" By answering these questions it is hoped the reader can gain an understanding of the current capabilities and limitations of automated speech systems. With this background, the issues of applying speech systems to military problems is undertaken. Because the use of speech systems is now so widespread the treatment here is not exhaustive, but instead is intended to be representative of much of the current research. This paper is a general survey and cannot comprehensively cover all of the topics presented. The extensive bibliography provides references for more detailed information for the interested reader.

## 2. What are Voice-Interactive Systems?

Potential users in industry, the home and the military are starting to become excited about the possibilities of voice interaction with machines. Speech technology has recently received considerable publicity as new applications are discovered (Doddington, G. R., and Schalk, T. B., 1981; Simmons, E. J., 1979; Levinson, S. E. and Liberman, M. Y., 1981). In this section the history of speech technology is traced from the pioneering days in the fifties to the present, and the advantages and disadvantages of speech are discussed. Prior to this discussion, it would be useful to review some terminology. Figure 1 lists some of the terms presently used in voice processing technology. Voice-Interactive systems includes both Automatic Speech Recognition (ASR) and systems for speech synthesis. "Speech Recognition" of a human speaker who utters single words or short sequences of words. "Speech Synthesis", sometimes called voice responses or voice output, refers to a machine which can generate a human or human-like vocal response. "Speech Understanding" is sometimes used to refer to machines which can not only recognize complete sentences (as opposed to a single word or short sequence of words) but somehow interpret the meaning of the sentence as well. "Speaker identification" (or speaker recognition) is used to describe a machine which has the ability to determine who is speaking, rather than what is said. There are other useful actions based on speech that can be performed by machines including: the capability to recognize what language is being spoken (language identification), the ability to remove noise or interference from speech (speech enhancement), the capability to compress the bandwidth necessary to transmit digital speech (vocoders), the capability to detect abnormal physical or psychological conditions of the speaker (stress analysis), and others. The useful functions being performed by these systems are accomplished through the theory and practice of speech processing technology. How this technology has evolved is the topic of the next section.

### 2.1 The Evolution of Speech Processing Technology

Speech processing technology has been based, to a large degree, on early research in such areas as experimental phonetics, the physiology of the human vocal apparatus and auditory system, human perception of speech, and especially acoustics, and phonetics. This basic research provided much of the fundamental knowledge which is required to some extent in almost all speech processing systems. One of the first applications of this knowledge to speech processing was reported in 1952 (Davis, K. H., et al, 1952) with the demonstration of a successful speech recognition system which could recognize the digits spoken from one talker. One of the first electronic speech synthesis systems was produced even earlier, in 1939, by researchers at Bell Labs (Dudley, H.; Riesz, R. R.; Watkins, S. A.; 1939). This system was called the Voder and was also a forerunner of the modern vocoder.

Important milestones in the development of speech processing technology occurred in 1956 and 1959, with the first efforts to incorporate linguistic information (Wrien, J. and Stubbs, H. L.; 1956) and the use of a general digital computer (Forgie, J. W. and Forgie, C. D.; 1959) respectively. One of the first speech recognition systems which could recognize continuous speech was developed in 1969 and could accommodate a highly constrained vocabulary of 16 words (Vicens, P. J.; 1969). Much of this early work assumed all of the information required to do recognition could he extracted from the spectral envelope of the acoustic speech wave. This resulted in the development of many approaches for spectral analysis of speech, and with these analysis approaches came very sophisticated mathematical techniques for manipulating the acoustic speech parameters. Some of these mathematical techniques are: linear predictive coding (LPC), dynamic programming, the Fast Fourier Transform (FFT), and homomorphic filtering, among others. Speech recognition techniques which are concerned solely with the manipulation of the acoustic waveform are sometimes referred to as mathematical, pattern matching or statistical approaches. However, an alternate approach was soon to receive considerable attention.

Prior to 1970 most of the work in ASR was concerned with recognizers which could only deal with a very limited vocabulary (typically 50 words or less), spoken in a discrete manner, for a talker who had previously used the device. In fact, many of these recognizers worked with high accuracy in the laboratory, and a group of researchers at RCA were encouraged enough to leave and form their own ASR company, Threshold Technology Inc., in 1970. However, the field of speech recognition was sharply criticized in a letter written by John Pierce, a highly respected scientist at Bell Laboratories (Pierce, J.; 1969). The letter accused researchers working in speech recognition of failing to appreciate the difficulty of their task. Although the letter seemed to put a temporary damper on the enthusiasm for speech recognition, the Advanced Research Projects Agency (ARPA) nonetheless funded a large, 5 year effort in the field in 1971. The ARPA project addressed the problem of speech understanding, rather than speech recognition, and had a number of ambitious technical goals. There is considerable debate even now as to the progress made in the project (Klatt, D. H., 1977; and Neuberg, E. P., 1975). What is noteworthy is that the approach taken in the project can be considered from the perspective of artificial intelligence (AI). Unlike the mathematical approach, AI presumes that a perfect extraction of phonetic features in speech is not necessary (or maybe even possible) because errors made in this extraction phase can be compensated by knowledge obtained from so-called "higher sources." This higher order knowledge includes syntax, semantics and the pragmatics of discourse. It remains to be seen which approach will be more successful. Perhaps a combination of approaches, along with greater computing power, will solve many problems. Most agree that increasing the knowledge of the human speech process is required before the effectiveness of speech systems match the expectations of potential users.

The last half of the seventies has seen increased attention given to the application of ASR technology to a variety of real-world problems, with less emphasis being given to more fundamental research. There are still many of these fundamental research problems which remain to be solved, as shall be discussed below. A number of companies, large and small, now have speech recognition products commercially available. An increasing number of speech synthesis products are also becoming available in the marketplace.

The state-of-the-art in speech recognition and synthesis will now be addressed. Practical ASR is restricted to discrete-utterance, limited vocabulary, and speaker dependent recognition of high quality speech. The accuracy of such systems are dependent on a variety of factors but accuracies near 100% in the laboratory and less than 90% in field tests are typical. Synthesis systems are generally of three types: 1. Those which do a simple encoding of the speech signal. An example is a simple digitization of real speech. The synthesis would then be accomplished by digital-to-analog conversion; 2. A complex encoding of the speech signal. An example of this type is linear predictive coding. Speech is encoded and then stored to form pre-recorded messages which are synthesized by doing the inverse of the encoding process; 3. Synthesis-by-rule systems which require very little storage of actual speech, but instead accept as input typed commands. The commands are interpreted and a basic set of speech sounds (phonemes) are strung together and modified by a complex sequence of rules. There are three main trade-offs associated with speech synthesis systems. These are: speech quality, memory requirements and message flexibility. The chart below summarizes these trade-offs for the three types of synthesis:

| SYNTHESIS TECHNIQUES | QUALITY | MEMORY | FLEXIBILITY |
|---|---|---|---|
| 1. Simple Encoding | High | Greatest | Moderate |
| 2. Complex Encoding | Moderate | Moderate | Low |
| 3. Synthesis-By-Rule | Low | Low | High |

Currently, there are more than 44 companies producing speech synthesis products (Wong, D., 1981). Excellent summaries of speech processing technology evolution can be found among the references (Denes, P. B. 1975; Hyde, S. R., 1972; Reddy, D. R., 1976; Lea, W. A., 1979).

## 2.2 Voice-Interactive System Defined

The previous discussion highlighted speech technology, not voice-interactive systems. The term "voice-interactive system" emphasizes the interfacing of a human and a machine that is of interest. The "voice" part of a voice-interactive system can mean either a human voice talking to a machine, or vice versa. Since a human is involved, it is not only speech technology that is of concern, but the psychology and physiology of the man-machine interaction. Researchers involved with speech processing are typically electrical engineers, computer scientists or mathematicians. Those involved with voice-interactive systems have a more behavioral science orientation, and include experimental psychologists and human factors engineers.

What, then, is meant by the term "voice-interactive system"? Conceivably it could mean any system involving humans and machines, with speech as the mode of communications. Thus, a digital voice communications system could qualify as a voice-interactive system under this definition. However, this is not what is usually meant by the term.

A Voice-Interactive System is defined as the interface between a cooperative human and a machine, which involves the recognition, understanding or synthesis of speech, to accomplish a task of command, control or communications, and which involves feedback from the listener to the speaker. With this definition, the digital communications system no longer qualifies, because the system provides an interface between a human and another human, not a machine. Likewise, speaker identification and language identification systems do not qualify as Voice-Interactive Systems because they involve non-cooperative speakers and no feedback from the speaker (human) and the listener (machine). Figure 2 shows in a very simple way a voice-interactive system: The box labeled "Speech I/O Subsystem" is some type of speech processing technology. Suppose the voice-interactive system was one in which a human pilot can control certain cockpit functions, and in addition can receive audio warning messages. The diagram of Fig. 2 can then be drawn more specifically as shown in Fig. 3. The interaction diagrammed in Fig. 3 is fairly complex and is intended to show relationships among all the elements involved, not any particular system. The pilot and speech I/O sub-system are both listeners and speakers. The pilot controls certain cockpit functions (which have not been specified) by speaking utterances into an ASR device. The controller of the voice-interactive system interprets the results of the recognition and responds with the appropriate controlling actions to the aircraft. Suppose an emergency situation arose of which the pilot was unaware. Presumably, the aircraft would signal this to the controller which would respond with the appropriate synthesized warning message. The pilot would then take corrective action which may require him to use manual controls, and the warning message would be subsequently halted. In this case there is feedback in both directions between man and machine.

## 2.3 Advantages of Speech Communications

There are good reasons why people might wish to use speech to communicate with machines and many reports have detailed the relative advantages of speech communications (Lea, W. A., 1968; Lea, W. A., 1979; Martin, T. B., 1976). However, there is relatively little empirical evidence which demonstrates the value of speech over other modes of communications, command or control. What empirical evidence does exist seems encouraging. In a famous experimental run at Johns Hopkins University, teams of people interacting together to solve problems solved them much faster using voice when contrasted to other modes of communication (Ochsman, R. B. and Chapanis, A., 1974). Other studies indicating the advantage in terms of speed and accuracy of voice over other modes of communications for certain tasks have been reported. (Welch, J. R., 1977; Harris S., Owens J., and North R., 1979; Skriver, C., 1979; Wherry, R., 1974; Poock, G. K.., 1980). Despite the lack of supporting data, a list of advantages shall be presented for speech communication in general, and in a later section for the application of speech in the cockpit environment. Many of these arguments for speech are of the "common sense" variety and there are undoubtedly others that could be added to them.

The most powerful reason for using speech is the fact that it is man's most natural form of communications and does not require special training to learn. A second strong argument for voice is that it frees the hands and eyes for other tasks. Most of the other advantages follow directly from these two. Figure 4 shows a list of advantages of speech communication. The list has been divided into three sections: engineering, psychological and physiological.

## 2.4 Disadvantages of Speech Communications

The disadvantages of speech communications should be considered carefully. It is important to make a distinction between the drawbacks of speech communications in general and the limitations of current speech technology. The former is relevant in speculating about the long-range possibilities of speech, and the latter is relevant to near-term concerns. The general disadvantages of speech communications are shown in Fig. 5, and those associated with cockpit environments are discussed in a later section. The disadvantages of speech communications involve mainly the effects of a hostile environment on the speech signal directly, or indirectly by a change of the physical or emotional state of the speaker.

## 3.0 Issues of Cockpit Applications of Voice-Interactive Systems

The use of voice-interactive systems offers the potential for solving critical man-machine problems in the aircraft cockpit. These problems are severe in military aircraft, and especially in aircraft capable of high performance. In these aircraft, crew members are often forced to cope with a very high workload, caused by inefficient crew member stations, poor assignment of operator tasks, and an overwhelming number of displays and indicators. In summary, the human operator is overwhelmed with too much information and has too many visual/manual tasks to perform. There has been recent attention placed on using voice-interactive systems in the cockpit to reduce the operator workload problem and solve other man-machine problems. All three services in the United States, many NATO countries and considerable industrial effort has addressed the application of speech technology (Lane, N. E., and Harris, S. D., 1980; Coler, [illegible] 1980; North, R. A., et al, 1980; Wicker, J. E., 1980; Harris, S., et al, 1980; [illegible] 1981, Reed, L., 1981). Because this is a tutorial paper, the discussion [illegible] summary of the relevant issues and not a detailed technical [illegible]

### [illegible] and Disadvantages of Voice Interactive Systems in the Cockpit

[illegible] future high performance attack/fighter aircraft may exceed [illegible] interface capabilities. Hence, a real-time voice interactive [illegible] potential solution as a method of augmenting current control/display [illegible] result of this realization, a number of airframe manufacturers have [illegible] and experimental design in interactive voice command/feedback [illegible] seat fighter aircraft.

[illegible] experienced military pilots were questioned regarding the idea of [illegible] for fighter aircraft. In one study, (Ruth, J. C. et. al., [illegible] were asked to rate, on a scale of zero to ten, the use of voice command [illegible] for high performance military aircraft; zero, of course, represented [illegible] the whole idea" and, ten represented "sounds great." These pilots were opposed [illegible] command for important decision making functions such as firing weapons, [illegible] eject and control trim. However, they were in favor of mode selection [illegible] functions such as radio channel selection, TACAN ILS, radar, bomb/NAV mode selection and IFF/transponder setup. Some early results indicate that voice command functions can't be directly substituted for control/display interfaces in a fighter aircraft.

In addition to the above, the military aircraft environment levies a number of additional requirements on automatic recognition subsystems (voice recognition devices). Some of these requirements are listed in Table 1.

TABLE 1

Oxygen Mask
Microphone
Physical Stress
Emotional Stress
Vibrational Effects
Complexity
Syntax

High Ambient Background Noise
Preselected Vocabulary
Non-robust words
Human Error
Overall reliability
Cost, Size/Weight

The requirements listed in Table 1 cause specific technical problems such as word boundary detection, memory requirements, small space, noise stripping and voice inconsistencies.

Questions arise as to whether training should be done with pilots wearing oxygen masks under actual flight conditions (different G-levels, engine power levels, canopy on-off). Types of signal input to system must include the affects of regulation, inhaling, exhaling, etc. Results have shown that because breath and background noise cause drop offs at the ends of words, an end point detector based on energy level can't be used, hence more sophisticated automated end point detection is required.

## 3.2 Speech Synthesis in the cockpit

Military aircraft applications of speech synthesis systems have also been investigated especially for caution and warning messages. Some of these applications are listed in Table 2.

Table 2

Applications of Speech Synthesis Systems

Voice Warning
Time-to-Go Countdown
Fault List Feedback
Way Point Announcement
Voice Response for Specific Request System Data
Audio Feedback to Voice Commands

A few problem areas resulting from speech synthesis in a high performance aircraft are specific message selection and corresponding voice quality. These synthesis systems must be aware of pilot safety, sound level variation for different noise levels, potential interference with other audio communications, cognitive and attentional demands.

It can be said there is general agreement that voice command, control and synthesis systems can provide the military aircrews with a useful adjunct to conventional control and display interfaces and provide warning and status data via speech synthesis. However, in order to apply this technology, many behavioral and human factors problems must be solved as well as some very difficult technical speech recognition issues. Clearly, the question of complexity and overall reliability of a voice interactive system in an aircraft environment must be addressed.

4. Other Military Applications of Voice Processing Systems.
(Beek, B., et.al., 1977, 1978, 1982)

1. Digital Narrowband Communications Systems.

Air Force tactical communications are being required to operate in increasingly difficult and hostile situations. Requirements are being levied on spread spectrum communications systems to provide increased communications capacity, multiple access, and tactical conferencing. Higher degrees of jam resistance and a lower probability of intercept are required in the already overcrowded, dynamic channel, and rapidly changing signal and interference environment. (See Fig. 6)

These increased requirements stress using existing frequency hopped, pseudo noise, voice FH/PN spread spectrum communications systems and ordinary HF and VHF radio systems. Systems being considered in exploratory development address these demands aggressively with a combination of Speech Processing, Adaptive Speech Processing, Adaptive Signal Processing, and VHSIC type microcircuitry.

Presently, the standard method of voice digitization being used is 16 kilobits CVSD. For advanced applications 2400 bit/sec LPC based systems are completing the developmental process. The 2400 bit/sec LPC system provides a factor of 6.66 reduction in input data rates which of itself would allow that many more channels in a communications bandwidth or a factor of 8 db increase in processing gain.

Recent research has produced sufficiently intelligible demonstrations of advanced exploratory techniques which are capable of digitizing continuous speech and retaining a degree of speaker recognition at rates down to 400 bits/sec. An additional factor of 6 increase over "standard" LPC is obtained providing that many more channels or another 8 db increase in processing gain. Total gain that can be achieved here is 36 additional channels or 16 db increase in processing gain for the potential AJ systems.

Isolated word recognition can further reduce the transmission rate to about 80 bits/sec for the limited vocabulary case. Basic research is presently underway applying artificial intelligence methods to achieve continuous speech recognition with less limited vocabularies. Fig 7 shows a simplified listing of the state-of-the-art of voice digitization system and limitations. Another factor of 5 is obtained here, with a corresponding 7 db increase in processing gain. Total gain that can be achieved is 180 additional channels or a processing gain advantage of 22 db.

Unfortunately, it becomes correspondingly more difficult to realize these gains in practice. For example, the total delay incurred in going through the speech processor/synthesizer combination grows as the degree of sophistication of the processor increases. For the intermediate state of voice compression the challenge is to achieve delays below 100 msec. For word recognition systems, delays of up to 250 msec. are not acceptable. Additionally, the longer integration times required for the longer bit times and the higher anti-jam requirements in many cases exceed the coherency of the channel. Electromagnetic compatibility demands consideration of shorter transmissions implying the use of lower duty cycle transmissions. The resulting loss in transmitted energy per bit requires an increase in the peak power of the transmitted signal not desirable for low detectability considerations. The incoherency of the channel taken together with the shorter pulse times will necessitate the use of incoherent combining techniques incurring additional losses. Finally, cockpit noise and speech distortion can increase the difficulty in successfully digitizing the speech information.

Fortunately, the application of advanced signal processing techniques can minimize the losses incurred. Adaptive signal processing and signal encoding techniques are being applied by current RADC development programs to achieve even more jam resistance and interference suppression than can be achieved by input data compression .echniques taken alone. Noise suppression and speech analysis efforts are showing great promise in solving the practical cockpit speech input problem. Additionally, basic research is being conducted to make practical the use of word recognition techniques for applications where the speech processing delay does not pose an unacceptable factor in the communications system design.

The RADC in-house program (HF Terminal with ECCM Modem, Speech Recognition/Synthesis) demonstrated a combination of techniques which provide anti-jam (AJ) voice communication over radio channels whose bandwidth ordinarily supports only conventional non-AJ voice (Beek, B., 1982). Moreover, this combination also provides enhanced .eliability under noisy (but unjammed) channel conditions. The voice source encoding employs special codes to represent phrases and in some cases, sentences, and thus provides a certain significant amount of data compression. This type of system will narrow the bandwidth requirements for voice communication to approximately 80 Hz and will provide 15.7db anti-jam margin. This compares very favorably with analog systems that require a bandwidth of 3000 Hz and has no anti-jam margin. Low data rate systems have the disadvantages of vocabulary size restriction, word rate restrictions, and loss of speaker identity, but the advantage of increased intelligibility may outweigh the disadvantages for certain applications. As connected speech recognition systems are developed, vocabulary size and word rate restrictions can be minimized.

## 5.0 Automatic Speaker Verification/Identification

Speaker Verification. The objective of this program is to develop automated methods of identity verification for the purpose of providing controlled access to secure areas. (See Fig. 9) For many years, RADC has supported the development of a method of entry control using speech as the personal attribute. The Automatic Speaker Verification (ASV) System has proved to be highly reliable (over 99% accurate) at verifying individuals' identity and detecting imposters.

An Advanced Development Automatic Speaker Verification System was fabricated, tested, and evaluated for entry control using a person's voice as a personal attribute for secure access control. Under this effort, algorithms were implemented on three TI 900 minicomputers, which were operationally tested for six months at the entrance of the Semiconductor building at Texas Instruments, Dallas Texas. A total of 286 users (200 men and 86 women) provided 13,539 accesses. A Type I error rate (true speaker rejection) of less than 1.0% was achieved. Off-line tests on casual impostors provided a Type II error rate (impostor acceptance) of less than 1.0% with a confidence level greater than 90 percent.

A study of speakers using an LPC-based prediction residual was also investigated under this effort. This study provided a magnitude of improvement in performance which exceeds the goals of this effort. Future work in this area is to implement an LPC-based speaker verification system.

Speaker Identification. This problem is similar to the speaker verification problem except no prior identity claim is made by the unknown speaker. Speaker identification is the harder problem for several reasons (See Fig. 10):

a. The speaker may be uncooperative;

b. The quality of the communications channel may be poor;

c. There is no control over the spoken text by the communications analyst;

d. The unknown speaker may or may not be a member of an original set of speakers; and

e. The recording and/or channel conditions may be different for speech collected for reference and test samples.

An exploratory development program to do speaker identification was recently concluded (See Fig. 11). The goals of the effort were to recognize any one of 30 unknown male talkers, using as little as ten seconds of reference and test speech data, in real-time as shown in Fig. 12. All goals of the effort were met or exceeded. These encouraging results were achieved by use of an algorithm originally developed by Markel, which uses ten Linear Prediction Codes (LPC) coefficients that are averaged over the entire recognition period. A follow-on effort is planned which will attempt to improve human factors aspects of the speaker identifications system and to improve recognition accuracy under noisy (10db or less SNR) channel conditions.

## 6. Speech Enhancement

The use of Automatic Speech Recognition (ASR) to relieve flight crew workload and to provide narrowband communications for airborne operations is highly desirable. Unfortunately no ASR system exists that can cope with the harsh, noisy airborne environment. Current commercial ASR equipment has not been designed to operate in the airborne environment. For this reason a considerable amount of attention has been given to reducing the effects of the airborne environment on ASR.

There are many environmental effects that cause poor operation of an ASR system in the aircraft environment. Some of these effects are aircraft noise, breathing noise, operator stress, operator fatigue, effects of gravitational forces on operator's speech, etc. Although all of these environmental effects must be reduced, much attention has been given to reducing the acoustic noise generated by the aircraft. The level and characteristics of this noise can vary considerably, depending on such conditions as type of aircraft, location of the ASR microphone, facemask or no mask operations, and status of aircraft.

The areas of concentration in reducing the effects of this noise have been in the development of more robust recognition algorithms and the development of techniques to reduce the acoustic noise before recognition processing begins. One area which has generated some interest for removing aircraft noise has been the area of speech enhancement. Some of the problems with these techniques have been high spectral distortion, limited noise adaptation, and distortion characteristics that vary with input signal noise level and spectral shape.

Rome Air Development Center (RADC) has been developing speech enhancement technology to improve the quality and intelligibility of speech signals that are masked and interfered with by communication channel noise. RADC's interest in speech enhancement is not only in improving the quality and intelligibility of speech signals for human listening and understanding but to improve speech signals for machine processing as well. Speech technology such as speaker identification, language recognition and keyword recognition being developed by RADC requires good quality signals in order to provide effective results. The development of automatic, real-time speech enhancement technology is therefore of high interest to RADC. This technology is required to improve the quality of degraded speech signals to an acceptable level for these systems.

Exploratory development work at RADC has led to the development of an Advanced Developmental Model enhancer called the Speech Enhancement Unit (SEU) (See Fig 13). This unit, which uses a high speed digital array processor in conjunction with time, frequency and root-cepstral algorithms, provides an on-line, real-time capability to remove frequently encountered communication channel interferences with minimum degradation to the speech signals. The types of interferences or noises removed can be classed into three groups; (1) impulse noises such as static and ignition noise, (2) narrowband noise which includes all tone-like noises, and (3) wideband random noise such as atmospheric and receiver electronic noises. Tests have shown that the SEU can reduce all of these types of noises simultaneously while improving both the quality and the intelligibility of the speech signal. The capability to remove both narrowband and wideband random noise without degrading the quality of the speech signal may make these speech enhancement techniques applicable to improving the performance of Automatic Speech Recognition (ASR) in the airborne environment. The SEU's ability to remove narrowband types of noises automatically and in real-time by as much as forty (40) decibels would allow the removal of such aircraft noises as power converter hums, periodic aircraft vibrational noises, aircraft compressor noises, and other rotational noises associated with the engine. Since the noise removal process causes little distortion to the speech signal and removes a minimum amount of the speech signal, this spectral noise removal process should remove all narrowband noises without having detrimental effects on the recognition accuracy of the ASR system.

The SEU's ability to remove wideband random noise automatically and in real-time may allow the removal of much of the unstationary noise generated by the aircraft. An example of the noise removal process is shown in Fig. 14. The wideband noise removal process is a root-cepstral process that can improve the signal-to-noise ratio of noisy communication channels as much as 12 to 14 decibels. An improvement of this amount in the signal received at the input of an ASR system could improve the performance of an ASR system vastly.

The wideband noise removal is a subtractive process that is accomplished in the spectrum of the square root of the amplitude spectrum. While this function is not the same as the cepstrum (the cepstrum is the spectrum of the log amplitude spectrum), it resembles the cepstrum and is referred to as the root-cepstrum. In this method of noise reduction the average root-cepstrum of the noise in the input signal is updated continually and subtracted from the root-cepstrum of the combined speech and noise. Because the random noise concentrates disproportionately more power in the low region of the root-cepstrum than does the speech, the subtracted reconstructed time signal produces an enhanced speech signal.

There are two reasons why this technique of wideband noise removal is encouraging for the successful removal of aircraft noise for ASR. First the noise removal technique used is independent of the spectral shape of the noise. This indicates that the enhancement unit should theoretically adjust to the aircraft noise. The second encouraging reason is that the enhancement transformation used, unlike the spectral subtraction methods which can cause high distortion, causes very little distortion to the speech signal which is important to the recognition accuracy of any ASR equipment.

The SEU's capability to reduce narrowband and wideband noise without causing distortion that is detrimental to the human listener (see Fig. 15) may be used to improve the recognition accuracy of ASR equipment in a noisy airborne environment. For this reason RADC is planning a series of carefully controlled tests. The tests will utilize two speech recognizers in conjunction with the SEU. The effects of various types of noise and on the recognition accuracy of these ASR systems will be determined with and without the enhancer. Preliminary results for an LPC Based Recognition System are shown in Fig. 16.

## 7. Voice Control & Data Entry Systems

A Voice Data Entry (VDE) system was designed for use in entering voice cartographic data to the Digital Landmass System (DLMS) data base. The first Voice Data Entry system was installed at the Defense Mapping Agency Hydrographic Center (DMAHC) (See Fig. 17). This allowed the user to enter depth information found on the map into a computer as shown in Fig. 18. This information was sorted along with the map coordinates of the particular depth readings. The vocabulary for this study included the digits plus a few control words. Results from this effort showed, for a limited vocabulary scenario where the operator had been sufficiently trained in system operation, that voice data entry was faster than a manual method of keyboard entry for both a skilled and unskilled operator. This effort also revealed an indepth study of error correction procedures, methods of system training, and operator familiarization procedures would be required in order to increase the efficiency of future Voice Data Entry Systems.

The second effort was the design and testing of a Voice Data Entry (VDE) system which would serve to input cartographic data to a computer. The system was installed at the Defense Mapping Agency Aerospace Center (DMAAC), for test and evaluation. The VDE system is intended for use in entering, by voice, cartographic data to the Digital Landmass System (DLMS) Data Base. The VDE system developed had the capability of recognizing up to 248 separate words in syntactic structures.

The two systems described are isolated utterance speaker dependent systems. For inputting a string of words, this requires a distinct pause between each word. Tests have shown that isolated word systems are three times slower, and more frustrating than normal voice data entry. This increases errors and further decreases the data entry speed. However, in many applications the emphasis is to input connected digits and isolated words or phrases. In these applications many of the functions/commands have been reduced to a set of digit codes well understood by the analysts.

Presently RADC is developing an Advanced Development Model (ADM) Voice Data Entry System to satisfy DMA's operational requirements for automated compilation of the Feature Analysis Data Table (FADT) for DLMS operation. This system will incorporate a limited vocabulary which may be entered in connected or normal speech, and an extended vocabulary which will be entered in an isolated speech mode.

RADC is also investigating voice interactive I/O algorithms to input a limited vocabulary spoken in continuous text into a computer with a voice synthesis feedback capability. The algorithms shall be capable of recognizing a 300 word, syntax independent vocabulary. The recognition shall be done in real time using a pretrained reference library.

Automatic Speech Data Entry Systems have application to many Air Force command, control and communication problems. However, the cost, size, weight, and power consumption of these devices must be reduced for many applications. RADC is currently looking at Very Large Scale Integration (VLSI) technology and microprocessor technology as a means of reducing cost, size, weight, and power consumption of VDE devices (See Fig. 19).

## 8. DOD and NATO Advisory Groups on Voice Technology

At the present time, two major military automatic speech recognition and technology groups are pursuing active technical coordination, data exchange and cooperative research projects. The first is the DOD approved Voice Technology for Systems Applications Sub-technical Advisory Group (VSTAG). The purpose of this VSTAG is to provide a forum for technical interaction between scientists and engineers at the bench level. Included as representatives to the VSTAG are members of the Air Force, Army, Navy, NASA, FAA, Post Office and NSA research laboratories that are engaged in speech processing applications. Table 3 lists the members of VSTAG.

The second is the NATO AC/243 Panel III Research Study Group (RSG)-10 for Speech Processing. The first meeting of RSG-10 was held in Paris, France in May 1978. Meetings are held twice a year and are rotated among the member nations. The technical objectives of RSG-10 are generally to review speech processing topics of military relevance in order to recommend specific research projects to be carried out cooperatively among the member nations. Member nations include Canada, France, Germany, Netherlands, United Kingdom and the United States. Table 4 is a list of of RSG-10 participants.

TABLE 3

Army

| | |
|---|---|
| ARI | Army Research Institute |
| ETL | Engineering Topographic Laboratory |
| AVRADA | Avionics Research Development Activity Human Engineering Lab<br>Communicative Technology Office |

Navy

| | |
|---|---|
| NAMRL | Naval Aerospace Medical Research Lab. |
| NADC | Naval Air Development Center |
| ONR | Office of Naval Research |
| NOSC | Naval Ocean Systems Center |
| NPGS | Naval Post Graduate School |
| NATC | Naval Air Test Center |
| NNMC | National Naval Medical Center |
| NWC | Naval Weapons Center |
| NASC | Naval Air Systems Command |
| NTEC | Naval Training Equipment Center |
| NPRDC | Navy Personnel R&D Center |

Air Force

| | |
|---|---|
| AFAMRL | Aero Medical Research Lab. |
| RADC | Rome Air Development Center |
| AFWAL | Air Force Wright Aeronautic Lab |
| AFIT | Air Force Institute of Technology |

Other Government Agencies

| | |
|---|---|
| IRS | Internal Revenue Service |
| USDA | Dept. of Agriculture |
| NBS | National Bureau of Standards |
| NSA | National Security Agency |
| OUSDRE | Office of the Under Secretary of Defense for Research Engineering |
| NASA | Ames Research Labs |
| | US Public Health Service |
| FAA | Federal Aviation Administration |

TABLE 4

| | | |
|---|---|---|
| Mr John S. Bridle | UK | (Chairman) |
| Dr M. Martin Taylor | Canada | (Secretary) |
| Dr Harmut Mutschler | FR Germany | (Delegate) |
| Mr Patrice DesVergnes | France | (Delegate) |
| Dr Harman J. Steeneken | Netherlands | (Delegate) |
| Mr Richard S. Vonusa | USA | (Delegate) |
| Dr Helmut Mangold | FR Germany | (Specialist) |
| Dr Joseph J. Mariani | France | (Specialist) |
| Dr Melvyn J. Hunt | Canada | (Specialist) |
| Dr Roger K. Moore | UK | (Specialist) |
| Dr Robert Breaux | USA | (Specialist) |
| Dr David Pallett | USA | (Specialist) |

## 9. Future Direction

Since its inception, research in automatic speech recognition (ASR) has progressed to the point where military application can be a reality. Man's most natural means of communication will be the future method of interaction with man's machine. Progress has been slow but steady and excellent success has been demonstrated on isolated word recognition devices and speech synthesis devices to make them practicable for military use. This has increased the interaction among scientists of various disciplines including interchanges - interaction in acoustic-phonetics, linguistics, signal processing, etc. In fact, as we have seen, international participation in the solution of numerous ASR problems is at hand.

However, although we have come a long way we still have a long way to go. Presently, we are too strongly focused on applications to extend the minimal support given to a number of fundamental issues. In fact, before ASR can even approach human performance, we still need significant advances in acoustic-phonetics relationships and English phonology.

# BIBLIOGRAPHY

1. Doddington, G.R. and Schalk, T.B., 1981, "Speech Recognition Turning Theory to Practice", IEEE Spectrum, Pg 26
2. Simmons, E.J., 1979, Speech Recognition Technology", Computer Design
3. Levinson, A.E., and Liberman, M.Y., 1981, Speech Recognition by Computer", Scientific American, Vol 244, No. 4
4. Davis, K.H., Biddulph, R. and Balashek, S. 1952; Automatic Recognition of Spoken Digits", J. Acoust Soc. Am. 124(6) Pg 637
5. Dudley, H., Riesz, R.R., and Watkins, S. A., 1939, "A Synthetic Speaker", J. Franklin Inst. 227, Pg 739
6. Wiren, J. and Stubles, H.L., 1936, "Electronic Binary Selection Systems for Phoneme Classification", J. Acoust Soc. Am. 30(8)
7. Forgie, J.W. and Forgie, C.A., 1959, "Results Obtained from a Vowel Recognition Computer Program", J. Acoust Soc. Am 31(11)
8. Vicens, P. 1969, "Aspects of Speech Recognition by Computer", Computer Abstr. 13(11), No. 3098, Nov
9. Pierce, J., 1969, "Whither Speech Recognition", J. Acoust Soc. Am., Vol 46, Pg 1049
10. Klatt, D.H., 1977, "Review of the ARPA Speech Understanding Project", J. Acoust Soc. of Am., Vol. 62 Pg 1345
11. Neuberg, E.P., 1975, "Philosophies of Speech Recognition", Speech Recognition, R. Reddy, ed., Academic Press, Pg 83
12. Denes, P.B., 1975, "Speech Recognition Old and New Ideas", Speech Recognition, R. Reddy, ed., Academic Press, Pg 73
13. Hyde, S.R., 1972, "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature", Human Communication, A Unified View, E.E. David and P.B. Denes, eds., McGraw Hill, Pg 399
14. Reddy, D.R., 1976, "Speech Recognition by Machine: A Review", Proceedings of the IEEE, Vol 64 Pg 501
15. Lea, W.A., "Speech Recognition: Past, Present and Future", Trends in Speech Recognition, W.A. Lea, ed., Prentice Hall, 1979
16. Lea, W.A., Establishing the Value of Voice Communications with Computers", IEEE Trans on Audio and Electroacoustics, Vol Av-16, No 2, Jun 1968, Pg 184
17. Lea, W.A., 1979 "The Value of Speech Recognition Systems", Trends in Speech Recognition, W.A. Lea ed., Prentice Hall
18. Martin, T.B., 1978, "Practical Applications of Voice Input to Machines", Proceedings of the IEEE Vol 64, No. 4
19. Ochsman, R.B. and Chapanis A., 1974, "The effects of 10 Communications Modes on the Behavior of Teams During Co-operative Problem-Solving", International Journal Man-Machine Studies, Vol 6, Pg 579
20. Welch, J.R., 1977 "Automatic Data Entry Analysis", RADC-TR-306, Rome NY
21. Harris, A., Owens, J., and North, R., 1979, "Human Performance in Time Shared Verbal and Tracking Tasks", WAMRL-1259, Pensacola FL
22. Skiner, C., 1979, "Vocal and Manual Response Modes: Comparison using a Time-sharing Paradigim", RADC-79-127-60. Warminster PA
23. Wherry, R., 1976, "VRAS-A Voice Recognition and Synthesis System", S/D Digest of Technical Papers, Vol III, Los Angeles CA
24. Wong, A., 1981, "Alternate Coding and Synthesis Techniques", Course Notes, Making Silicon Talk and Listen, Signal Technology Inc., Santa Barbara CA, Part II, Pg 9
25. Lane, N.E., and Harris, S.A., 1980, "Conversation with Weapons Systems; Crewstation Applications of Interactive Voice Technology", Yearbook on Navy Manpower, Personnel Training, Research and Development
26. Werkowitz, E., 1980 "Ergonomic Considerations for the Cockpit Applications of Speech Generation Technology", Proceedings, Symposium on Voice-Interactive Systems: Applications and Payoffs, S. Harris, ed., Dallas TX, Pg 295
27. Drennen, T.G., 1980, "Voice Technology in Attack/Fighter Aircraft", Proceedings, Symposium of Voice-Interactive Systems: Applications and Payoffs, S. Harris, ed., Dallas TX, Pg 199
28. Coler, C.R., 1980, "Automatic Speech Recognition and Man-Computer Interaction Research at NASA-AMES Research Center", Proceedings, Symposium on Voice-Interactive Systems: applications and Payoffs, S. Harris, ed., Dallas TX, Pg 249
29. North, R.A., Mountford, S.J., Edman, T., and Guenther, K., 1980, "The use of voice entry as a possible means to reduce fighter pilot workload: A practical application of Voice-Interactive Systems", Proceedings, Symposium on Voice-Interactive Systems: Applications and Payoffs, S. Harris, ed., Dallas TX, Pg 251
30. Wicker, J.E., 1980, "Some Human Factors Aspects of Real-Time Voice Interactive System in the Single-Seat Fighter Aircraft", Proceedings, Symposium on Voice-Interactive Systems: Applications and Payoffs, S. Harris, ed., Dallas TX, Pg 265
31. Harris, S., Lane, N.E. and Curran, P.M., "The Navy Voice-Interactive Systems and Technology (VIST) Program", Proceedings, Symposium on Voice-Interactive Systems: Applications and Payoffs, S. Harris, ed., Dallas TX, Pg 311
32. Werkowitz, E. 1981, "Air Force Cockpit Utilization of Speech Technology", Proceedings, Conference on Voice-Interactive Avionics, S. Harris, Chairman, Warminster PA

33. Reed, L., 1981, "Voice-Interactive Systems Technology Avionics (VISTA)", Proceedings, Conference on Voice-Interactive Avionics, S. Harris, Chairman, Warminster PA

34. Beek, Neuberg and Hodge, "An Assessment of the Technology of Automatic Speech Recognition for Military Applications", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol ASSP-25 Pg 310-322, No. 4, Aug 1977

35. Beek, and Cupples, "Military Applications of Automatic Speech Recognition and Future Requirements", Voice Technology for Interactive Real-Time Command/Control Systems Applications, Cuman, Breaus and Huff (eds.), Proceedings, Symposium, NASA Ames Research Center, Moffett Field, Dec 1977

36. Beek, Broglie and Vonusa, "Voice Recognition as Related to Military Applications", Wescon Technical Papers, Western Electric Show and Convention, Los Angeles CA, 1978

37. Beek, Broglie and Vonusa, "Voice Data Entry for Cartographic Applications", Proceedings, American Congress on Surveying and Mapping, Oct 18-21, 1977

38. Beek, Manor and Woodard, "Speaker Authentication and Voice Data Entry", Proceedings, Twenty-first Symposium on Circuits and Systems, IEEE, Iowa, Aug 1978

39. Beek, B., "Overview of State-of-the-Art, R&D NATO Architecture & Possible Applications - Voice Processing Technology", NATO AGARD Prog. No 329, Blackpool, UK, April 1982

40. Vonusa, R. S.,Cupples, E. J., Steigerwald, Woodard, J. P.,and Nelson, J. T., "Application, Assessment and Enhancement of Speech Recognition for the Aircraft Environment", NATO AGARD Prog. No. 329, Blackpool, UK, Apr 82

41. Ruth, J. C., Goodwin, A. M., and Werkowitz, E. B., "Voice Interactive System Developemnt Program", NATO AGARD Prog. No 329, Blackpool, UK, Apr 82

42. Scott, "Word Recognition", RADC-TR-207, Sep 1978

43. Poock, G. K., "Experiments with Voice Input for Command and Control: Using Voice Input to Operate a Distributed Computer Network, NPS 55-80-016, Apr 1982

44. Scott, "Voice Data Entry for Cartographic Applications", Proceedings, Symposium on Voice Interactive Systems - Applications and Payoffs, Dallas TX, May 13-15, 1980

45. Welch, "Advanced Image Exploitation Aids", RADC-TR, to be published

46. Davis, Secrest and Cato, "Limited Vocabulary Continuous Word Recognition", RADC-TR-204, 1980

47. Doddington and Hydrick, "Speaker Verification II", RADC-TR-75-274, Nov 1975

48. Doddington, "Voice Identification for Entry Control", Proceedings, Symposium on Voice Interactive Systems: Applications and Payoffs, Dallas TX, May 13-15 1980

49. Markel, Oshika and Greg, "Long-Term Feature Averaging for Speaker Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol ASSP-25, No 4, Pg 330-337, Aug 1977

AUTOMATIC SPEECH RECOGNITION (ASR)

SPEECH SYNTHESIS/VOICE RESPONSE

SPEECH UNDERSTANDING

SPEAKER IDENTIFICATION/VERIFICATION

LANGUAGE IDENTIFICATION

SPEECH ENHANCEMENT

VOCODER

SPEECH DETECTION

SPEECH SPEED RATE CHANGE

Fig.1 Voice processing technology terminology

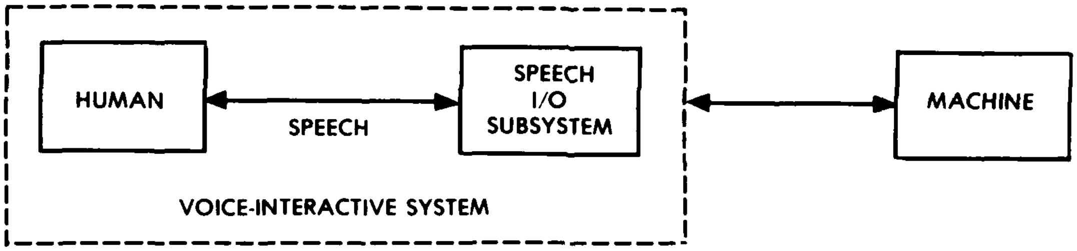


Fig.2 Diagram of a general voice-interactive system

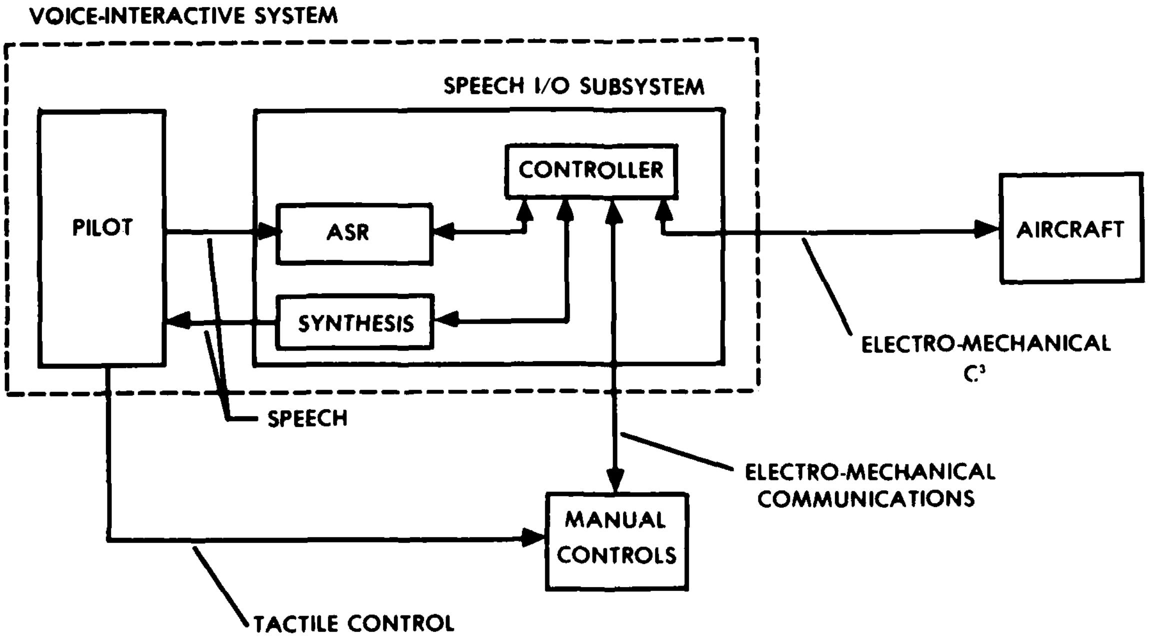


Fig.3 Voice-interactive system in cockpit setting

ENGINEERING

1. CAN BE FASTER THAN OTHER MODES OF COMMUNICATIONS
2. CAN BE MORE ACCURATE THAN OTHER COMMUNICATION MODES
3. COMPATIBLE WITH EXISTING COMMUNICATION SYSTEMS, E.G. TELEPHONES
4. CAN BE MORE ACCURATE AT TASKS CURRENTLY PERFORMED BY HUMANS, E.G. AUTOMATIC SPEAKER VERIFICATION vs IDENTITY VERIFICATION BY HUMAN VISUAL INSPECTION
5. CAN REDUCE MANPOWER REQUIREMENTS
6. CAN BE MOST COST-EFFECTIVE MAN-MACHINE INTERFACE

PSYCHOLOGICAL

1. MOST NATURAL FORM OF HUMAN COMMUNICATION
2. BEST FOR GROUP OR TEAM PROBLEM SOLVING
3. UNIVERSAL (OR NEARLY SO) AMONG HUMANS & REQUIRES NO TRAINING
4. CAN CONTAIN VALUABLE INFORMATION REGARDING EMOTIONAL STATE OF SPEAKER
5. CAN REDUCE VISUAL & MOTION INFORMATION OVERLOAD
6. CAN REDUCE VISUAL & MOTOR WORKLOAD
7. INCREASES IN VALUE PROPORTIONAL TO COMPLEXITY OF INFORMATION BEING PROCESSED
8. CAN REDUCE ERRORS FOR TASKS INVOLVING CONSIDERABLE COGNITIVE (AS OPPOSED TO PERCEPTUAL) EFFORT

PHYSIOLOGICAL

1. REQUIRES LESS EFFORT & MOTOR ACTIVITY THAN OTHER COMMUNICATION MODES
2. FREES EYES & HANDS & DOES NOT REQUIRE PHYSICAL CONTACT WITH TRANSDUCER
3. PERMITS MULTI-MODAL OPERATION
4. POSSIBLE EVEN IN DARKENED ENVIRONMENTS
5. OMNI-DIRECTIONAL & DOES NOT REQUIRE DIRECT LINE OF SIGHT
6. PERMITS CONSIDERABLE OPERATOR MOBILITY
7. CONTAINS INFORMATION ABOUT IDENTITY OF COMMUNICATOR
8. CONTAINS INFORMATION REGARDING PHYSICAL STATE OF THE COMMUNICATOR
9. SIMULTANEOUS COMMUNICATIONS WITH HUMANS & MACHINES

Fig.4 Advantages of speech communications

1. COMPETING ACOUSTIC SOURCES MAY INTERFERE WITH SPEECH. THESE INCLUDE NOISE, DISTORTION, & OTHER TALKERS
2. VARIETY OF PHYSICAL CONDITIONS CAN CHANGE ACOUSTIC CHARACTERISTICS OF SPEECH, INCLUDING VIBRATION, G-FORCES, & PHYSICAL ORIENTATION OF SPEAKER
3. HUMAN FATIGUE CAN RESULT FROM PROLONGED SPEAKING & FATIGUE MAY CHANGE SPEECH CHARACTERISTICS
4. PHYSICAL AILMENTS SUCH AS COLDS MAY CHANGE SPEECH CHARACTERISTICS
5. SPEECH IS NOT PRIVATE & MAY BE OBSERVED BY OTHERS
6. NO PERMANENT RECORD OF SPEECH UNLESS RECORDED EXPLICITLY (NOT TRUE OF TYPING)
7. PSYCHOLOGICAL CHANGES (STRESS FOR EXAMPLE) IN SPEAKER MAY CHANGE HIS SPEECH CHARACTERISTICS
8. MICROPHONES REQUIRED FOR SPEECH INPUT, ACOUSTIC SPEAKERS FOR SPEECH OUTPUT
9. SPEECH SYNTHESIS MAY INTERFERE WITH OTHER AURAL INDICATORS
10. SPEECH SYNTHESIS MORE SERIAL INFORMATION CHANNEL THAN VISUAL DISPLAYS & CAN BE SLOWER

Fig.5 Disadvantages of speech communications

- LIMITED CHANNEL CAPACITY
- BETTER NOISE IMMUNITY
- INCREASED JAM-RESISTANCE
- COST ADVANTAGES

Fig.6 Bandwidth reduction needed because:

UNDER 200 bps SYSTEMS

- SPEAKER DEPENDENT
- LIMITED VOCABULARY

200-400 bps SYSTEMS

- SPEAKER INDEPENDENT
- UNLIMITED VOCABULARY

Fig.7 State of the art

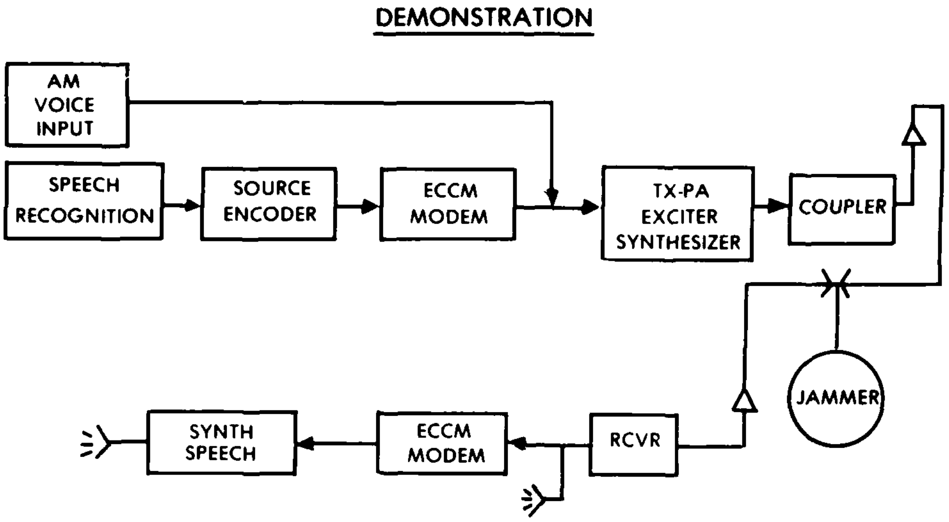


Fig.8 HF terminal with ECCM modem, speech recognition/synthesis

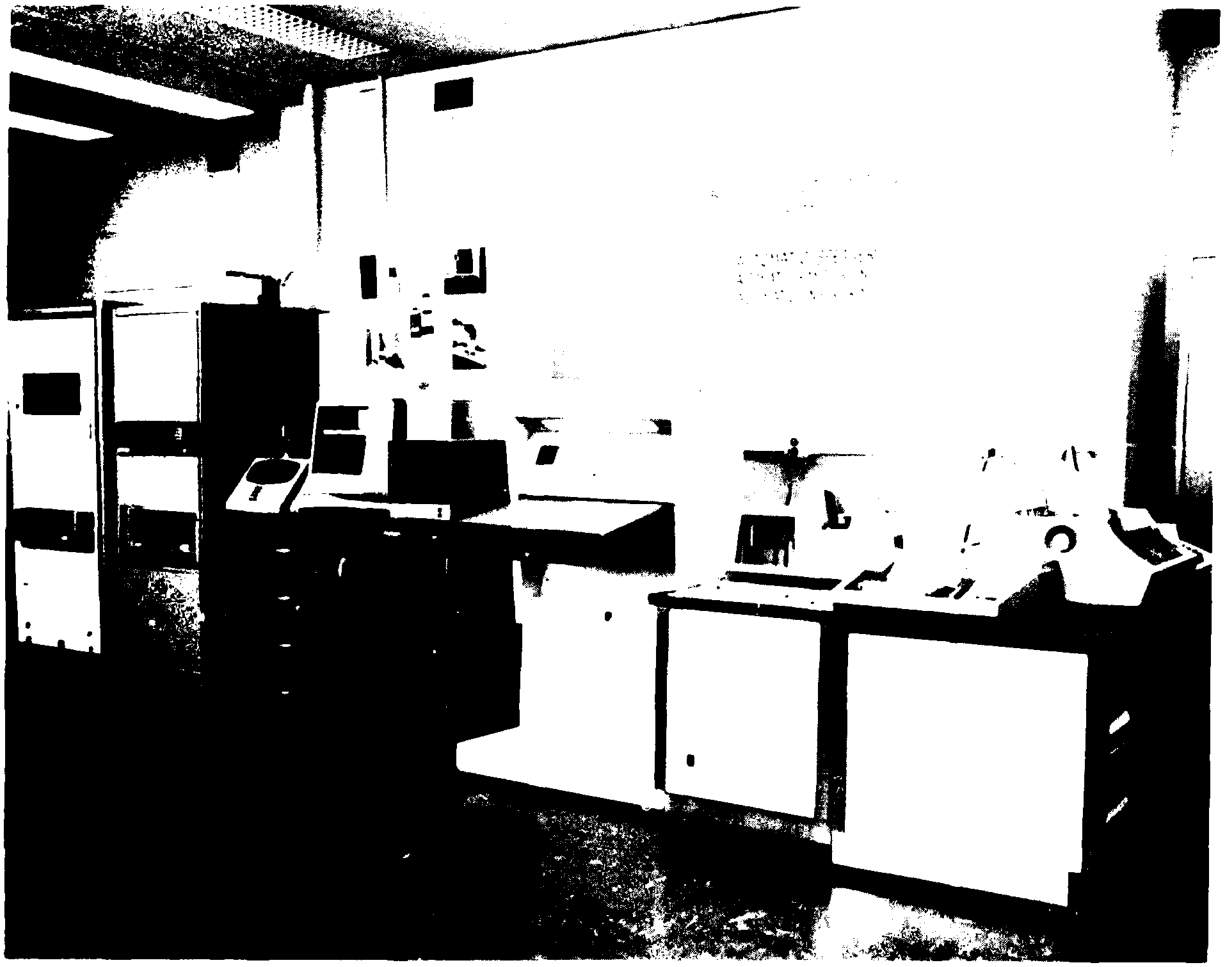


Fig.9 BISS in-house test facility

GENERAL CONSTRAINTS

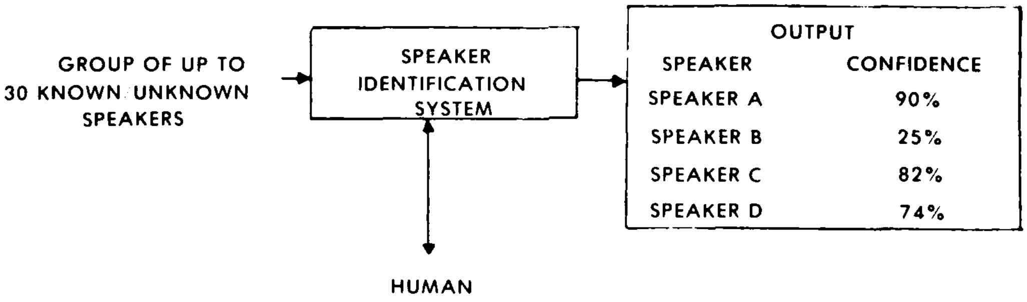
TEXT - INDEPENDENCE
UNCOOPERATIVE SPEAKERS
BAND - LIMITED & NOISY COMMUNICATIONS CHANNELS
CHANNEL CONDITIONS MAY VARY FOR SPEECH
COLLECTED FOR REFERENCE & TEST SAMPLES

OPERATIONAL CONSTRAINTS

MUST OPERATE ON - LINE & IN REAL - TIME
SPEECH SEGMENTS AVAILABLE FOR REFERENCES &
UNKNOWNS MAY BE VERY SHORT
MUST WORK RELIABLY FOR SEVERAL LANGUAGES

Fig.10 Speaker authentication problem

SYSTEM OPERATION



| SPEAKER | CONFIDENCE |
|---|---|
| SPEAKER A | 90% |
| SPEAKER B | 25% |
| SPEAKER C | 82% |
| SPEAKER D | 74% |

FEATURES:
- HUMAN HAS OPPORTUNITY TO OVERRIDE MACHINE'S DECISION
- HUMAN MAY UPDATE FILES WHEN HE DESIRES
- HUMAN MAY LISTEN TO MOST RECENT SPEECH DATA FOR ANY SPEAKER
- OPERATION IS REAL-TIME, ON-LINE, CONTINUOUSLY

Fig.11 Speaker identification

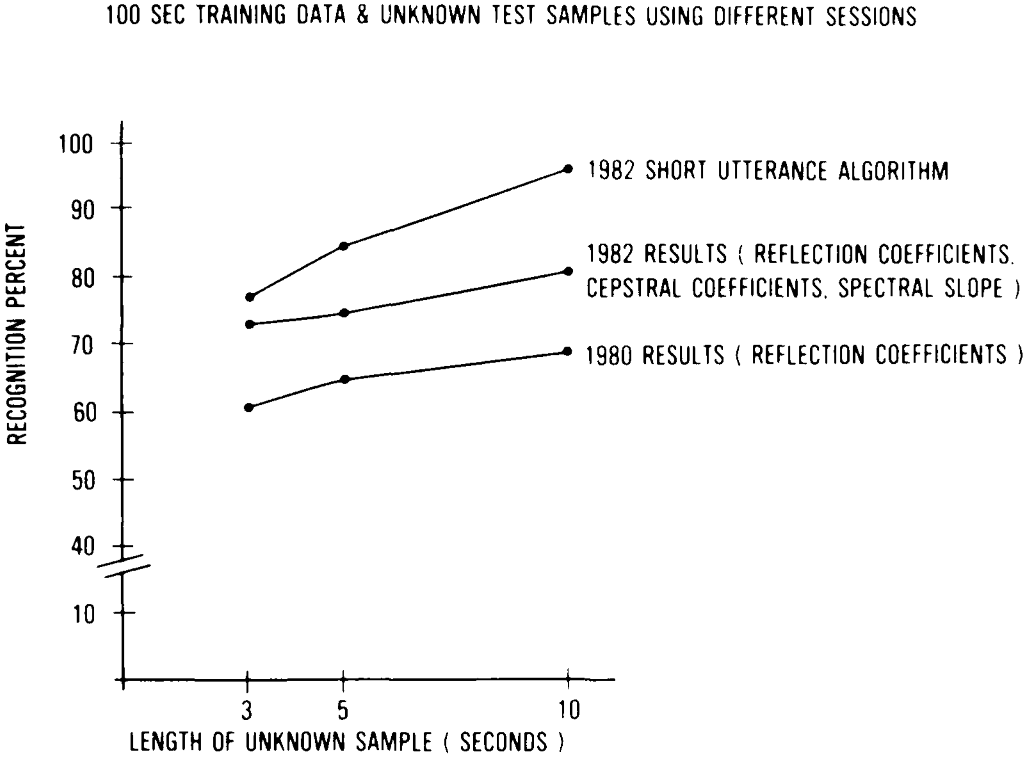


Fig.12 Automatic speaker recognition performance

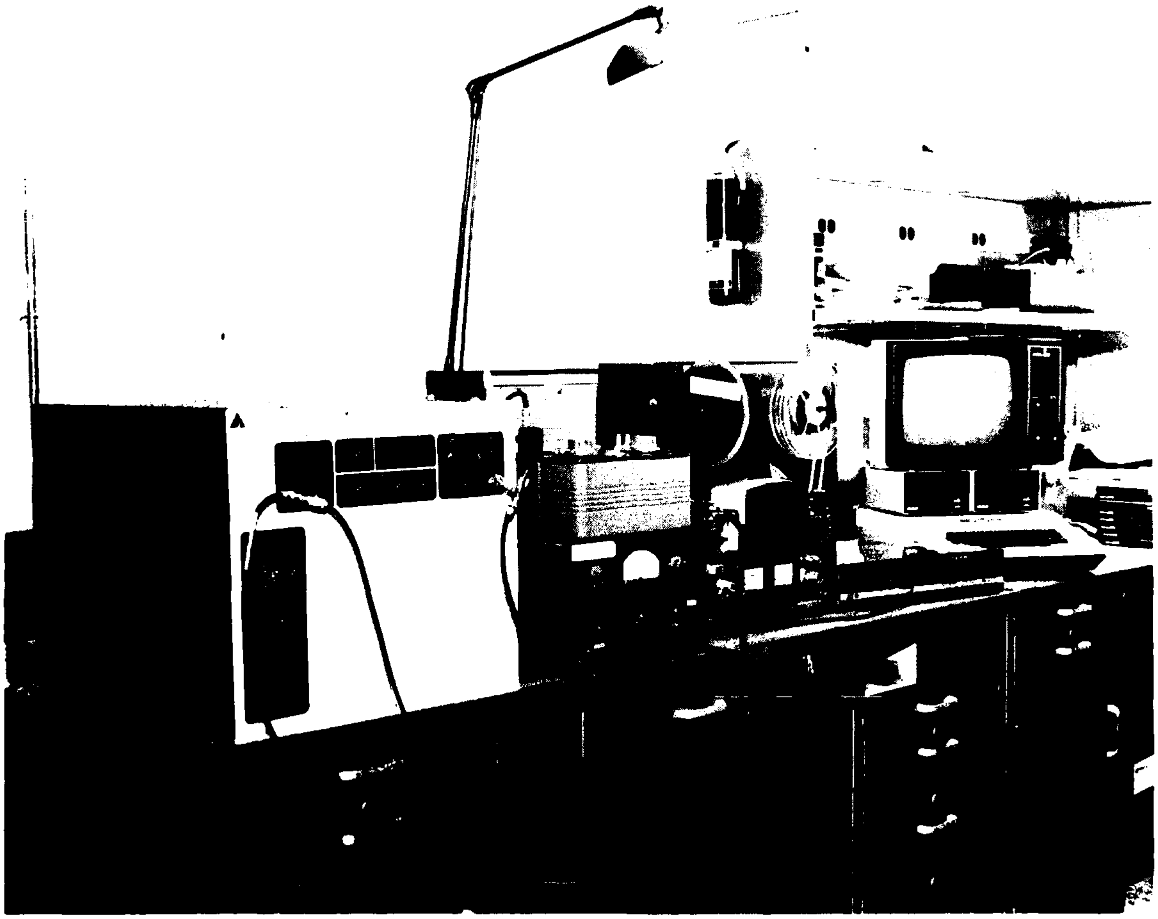


Fig.13 Speech enhancement unit

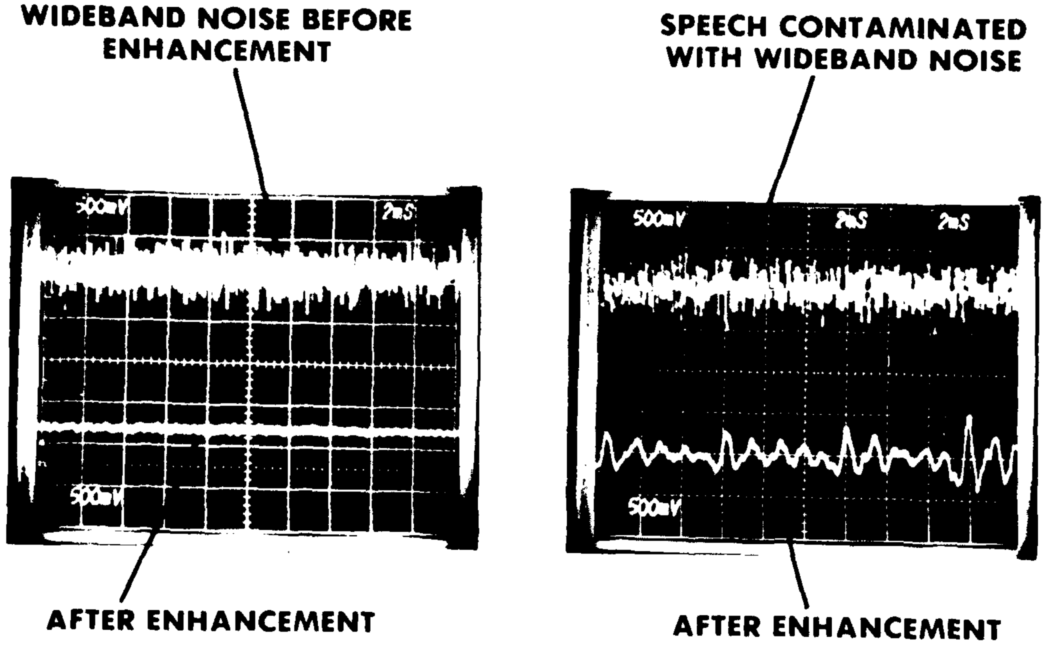

WIDEBAND NOISE BEFORE ENHANCEMENT
SPEECH CONTAMINATED WITH WIDEBAND NOISE
500mV
2mS
500mV
500mV
2mS
2mS
500mV
AFTER ENHANCEMENT
AFTER ENHANCEMENT

Fig.14 Wideband noise removal process

SIGNIFICANT IMPROVEMENT IN READABILITY OF MESSAGES

INTELLIGIBILITY IMPROVEMENT

SIGNIFICANT REDUCTION IN OPERATOR FATIGUE

AUTOMATIC PROCESS ADEQUATE

IMPROVEMENT IN EVENT RECOGNITION

Fig.15 Speech enhancement test results

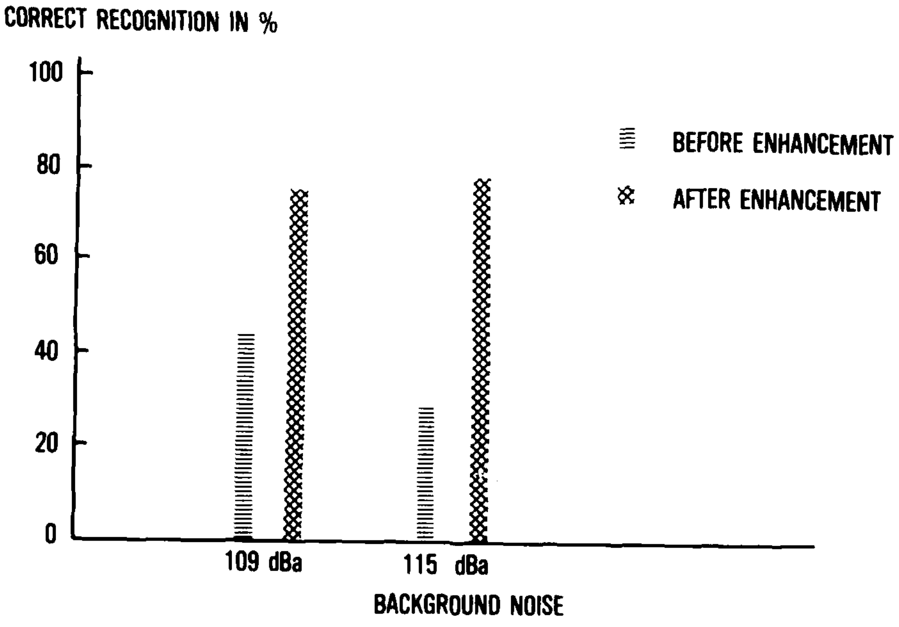


Fig.16 SEU/LPC based recognition performance

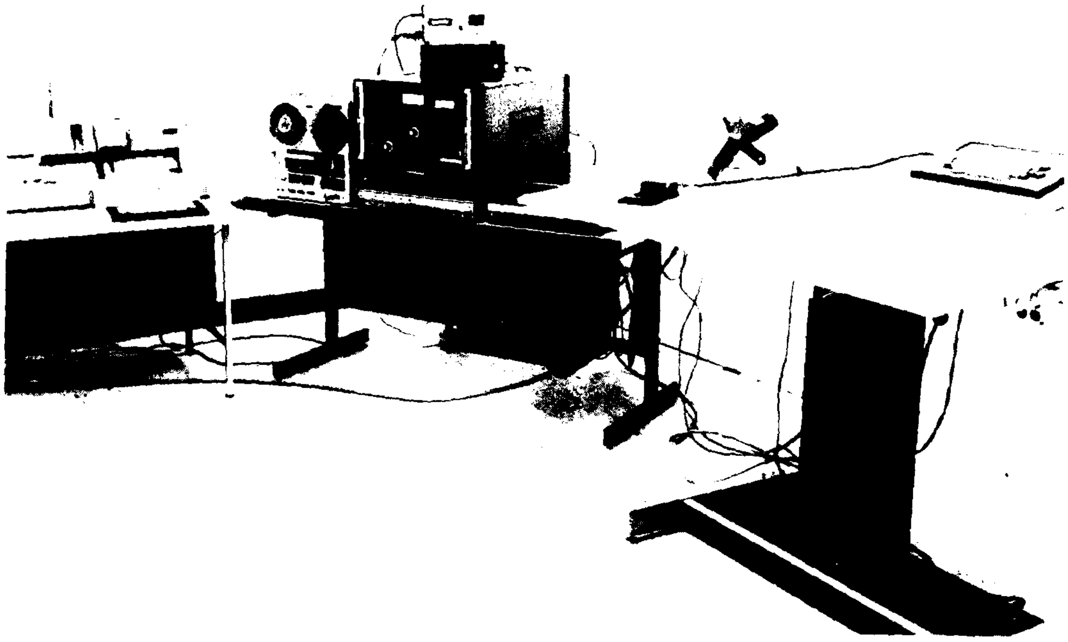


Fig.17 Voice data entry for DMA

Fig.18 Voice data entry testing

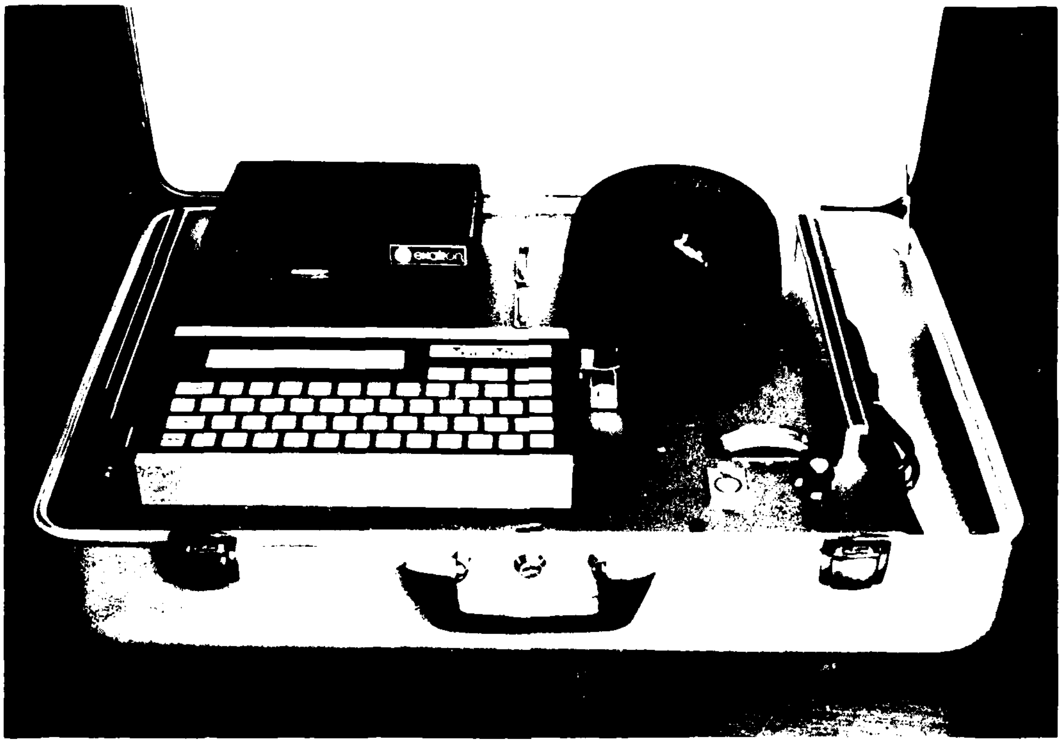
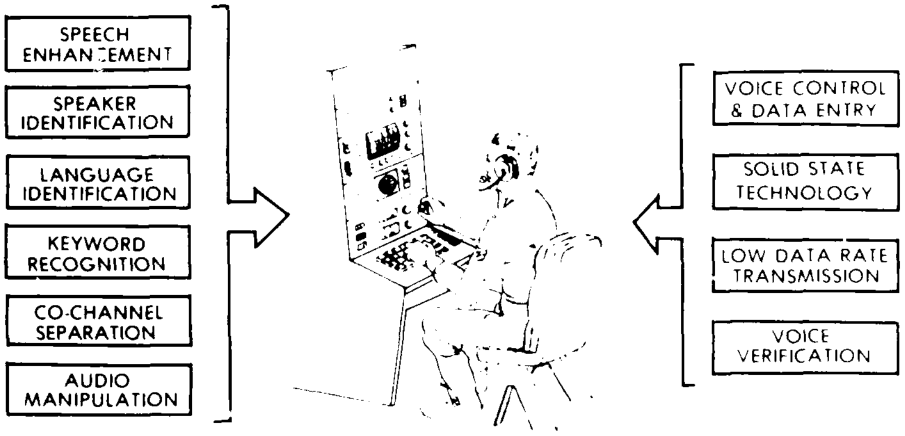


Fig.19 VLSI voice data entry system



GOALS

1 INCREASE OPERATOR PRODUCTIVITY
2 DECREASE OPERATOR FATIGUE
3 DECREASE REPORT PREPARATION TIME
4 SEMI-AUTOMATIC MONITORING
5 SEMI-AUTOMATIC TRANSCRIPTION
6 IMPROVE INTELLIGENCE OUTPUT

Fig.20 Tactical speech signal processing

# The Speech Signal

*Melvyn J. Hunt*

National Research Council of Canada
National Aeronautical Establishment
U66, Montreal Road
Ottawa, Ontario
K1A 0R6
Canada

## Summary

This talk is intended to provide an introduction to the speech signal with particular emphasis on the recognition of spoken messages. In an attempt to clarify its nature, speech is compared with two other kinds of signal. Some properties of words and phonemes are considered, and it is concluded that, unlike many artificial message-bearing signals, speech cannot be considered as a simple sequence of independent message units. A speaker adjusts the amount of information in his speech to suit his listener, and the listener carries out an active reconstruction of the message from the information available to him. Turning to recognition by machine, the use of syntactic constraints is first discussed, followed by a look at three kinds of approach to the analysis and representation of speech for recognition purposes. A brief account of speech production is provided in order to explain the motivation for production-based representations. This is followed by a look at how knowledge of auditory perception has been incorporated into recognition systems. Finally, some purely pragmatic approaches are discussed, and it is argued that success here generally correlates with simplicity.

## Introduction

This session is concerned with the nature of the speech signal itself: the signal that allows one human being to communicate to another whatever message he consciously chooses to express, with no external aids and usually with very little effort. As such, it is the session furthest removed from applications of speech technology. I am assuming that you, the audience or the readers of the proceedings, are mostly not speech specialists but rather people interested in how speech technology can be used. I am therefore not going to try to give you a comprehensive account of speech production or phonetics or linguistics. Instead, I want to put to you a few general ideas about the speech signal. My hope is that these ideas may provide a clearer picture of what people trying to make speech recognizers are up against, which recognition tasks are difficult and which relatively easy.

Before I start, I should mention a problem often faced by speech researchers in describing their work: if this lecture series were about some newly developed or newly discovered signal we could address an audience free of preconceptions, ready to accept whatever we had to tell them. But everyone can speak, and so everyone already has some strong subjective ideas about the speech signal. What is worse, most people can read, and their knowledge of the written representation generally has a strong effect on how they think of the spoken signal. I will come back to this point later. For the moment, perhaps you might try to forget that you can speak or read.

## What sort of a signal is speech?

In trying to answer this question, I think it is helpful to start off by considering two other types of signal that speech is sometimes grouped with. The first is a class of signals that are subjected to image processing. To be specific, let us choose a satellite image of a portion of the earth. Such an image has the obvious difference that it is two dimensional while the speech signal is effectively one dimensional. The more important difference, though, is that the satellite image is not a communication: it contains information but it does not contain a message. The very same image might be used to study the vegetation of an area or to try to spot missile silos, but presumably the image processing techniques appropriate for the one task would be quite different from those appropriate for the other. Thus, image processing tends to be a loose collection of techniques with diverse goals. Depending on which field we want to flatter, we can describe the automatic speech recognition problem as more limited or as more coherent than image processing.

The image processing problem we just discussed is rather like the problem in the speech field of determining the identity or the emotional state of a speaker from a speech sample, since we the receivers are deciding what information we want to derive from the signal rather than trying to extract the message being intentionally supplied by the speaker. The rest of this session, however, and indeed most of this whole series, is concerned with the problem of recognizing or efficiently transmitting the *intended* message, not the side information that may come with it.

The discussion that follows also excludes certain kinds of social communication such as "Hello, how are you?", where the speaker is not so much enquiring into the state of health of the listener as making a semi voluntary announcement of his feelings and relationship to the listener. This use of speech seems similar to the way in which a dog might bark a greeting at its master or a threat at an intruder. It is not what makes human speech special, and it is not of primary interest in communicating with machines.

The second kind of signal I would like to have you consider is a man made artificial communications signal. We could take as a specific example another optically derived signal like the output of a scanner reading product codes in a supermarket, but I think a better one is provided by an h.f. radio transmission carrying teleprinter text. In this example, there is quite clearly a message, and the message is laid out sequentially in time or space just like speech. The similarities to speech are obvious; the differences much less so, but they are nonetheless large and I want to take some time to look at them.

The artificial signals in our examples are composed of a sequence of units, the units being selected from a definite, known set that I want to call an *alphabet*. The units in a message are generally well separated from each other, and they do not interact. The decoding device usually has available to it in some form an *ideal*, undistorted representation of the alphabet, and decoding consists mainly of trying to identify the received units one by one using its built-in knowledge of the ideal forms.

Words

What is the equivalent of these units for the speech signal? I contend that there is no single exact equivalent. Perhaps the closest candidate is the word, but words differ in several major respects from our artificial units.

First of all - notwithstanding our prejudices from the *written form of language* - spoken words do not in general have gaps between them. Indeed, there are no consistent acoustic cues of any kind to word boundaries. What is more, not only are words not well separated from each other, they often interact at their boundaries. For instance, "bread board" is often pronounced in fluent English in a way that we might write as "breab board", and "this shop" as "thish shop". There is a more extreme example in French in the phenomenon of *liaison:* "ils ouvrent" *(they open)* sounds different from "il ouvre" *(he opens)* because we hear the "s" of *ils* in the first case. But it takes the initial vowel in *ouvrent* to bring the "s" to life: the corresponding expressions for closing - "ils ferment" and "il ferme" - do not have any distinction in their pronunciation.

Next, we know of no ideal reference forms of words: any normally pronounced version of a word is as good as any other, and no two productions will ever be exactly the same. In particular, words differ in their *prosodic* features (intonation, timing and loudness) depending on their function in a sentence. Even in such a prosaic utterance as a list of digits, the final digit differs markedly from the others, being typically 60% longer and having a falling intonation. When people try to generate synthetic sentences by recording words in isolation and playing them back unmodified in a sequence, the result is disastrous - each word is perfectly clear, but the sentence is almost impossible to follow.

Despite these problems with words, most of the more successful and practical connected speech recognizers have been word-based. As will be explained elsewhere, ways have been found to ignore prosodic differences and concentrate on the phonetic identity of words.

Phonemes

When I suggested words as the best equivalent of the artificial communication units, I imagine some of you were surprised that I did not choose *phonemes*. Such surprise would be understandable considering the number of popular articles on speech technology that talk about speech being made up of phonemes as though it were like laying out bricks in a line - just like the symbols in our teleprinter transmissions, in fact. Proponents of phonemes might also point out that the phoneme inventory (just over forty in English) is much more manageable more alphabet sized than the enormous inventory of words in a language. Some people might also be influenced by the way words are printed as a string of discrete context-independent letters. Despite all this, I want to suggest to you that phonemes bear very little resemblance to teleprinter symbols. If you would like a *writing analogue* for phoneme sequences, quite a good one is provided by hastily scribbled handwriting, in which individual letters are hard to isolate and depend for their form very much on the other letters around them.

A phoneme is defined as *the smallest unit of speech within a word that when changed results in a change in the meaning of the word.* Thus, the English word *tap* differs from the English word *cap* in the position of the tongue at the start of the two words. In *tap* the point of contact between the tongue and the roof of the mouth is just behind the upper teeth, while in *cap* it is at a point quite far back in the mouth. We can conclude that *cap* and *tap* must start with a different phoneme. We could have started with the tongue making contact in other places: it could have been directly behind the upper teeth like the "t" sound in eighth, or the tip of the tongue could have been curled back slightly like the "t" in *tree.* If we used either of these "t" sounds in our word *tap* we would *not* get a new word, we would simply have *tap* with a slightly non-standard pronunciation - we might not even notice that the word sounded odd if it occurred in fluent speech. Yet those same "t" sounds represent different phonemes for some other languages. For speakers of such languages (which include several languages spoken on the Indian subcontinent) the "t" variants presumably sound quite distinct. In the same way, the English "l" and "r" sounds in words like *lap* and *rap,* which sound quite different to English speakers, do not correspond to different phonemes in Japanese, so Japanese speakers have difficulty in making the distinction.

Phonemes, then, are not "speech sounds" in some absolute sense, they are a property of the way a language gets coded in sound, and their phonetic realization is frequently context dependent. Something interesting is happening in standard French right now the vowel sounds in the digits *deux* and *neuf* used to be different phonemes, that is to say, there existed at least one pair of words - *jeûne* and *jeune* are usually cited - that differed just by the fact that the first had the *deux* vowel in it and the second the *neuf* vowel. French speakers are increasingly using a new rule that says that the *deux* vowel can occur only at the end of a word and the *neuf* vowel only at a non-final position in a word. Thus the *jeûne/jeune* distinction is lost, and the two vowels have become context-dependent *allophones* of the same phoneme. French has lost a phoneme, but it has *not* lost a speech sound.

So far, we have established that phonemes do not correspond to a single speech sound, but perhaps we could say that it corresponds to a set of sounds. If by "sounds" we mean something we can hear and identify in isolation, the answer has to be no, or at least not always. The English word *do* is made up of two phonemes /d/ and /u/ (phonemes are conventionally written between oblique lines), but there is no way of pronouncing the /d/ without also pronouncing a vowel either before or after it. What is more, if we take a recording of *do* and listen to what happens as we shorten it by successively chopping off more and more of the vowel, we never get to hear a /d/ in isolation: when we have shortened it enough that we no longer hear the vowel, we no longer hear anything that we perceive as speech.

The picture of what a phoneme might be in acoustic terms gets even fuzzier when we start to ask about the acoustic features a listener might use to decide what phoneme sequence he is hearing. By using a speech synthesizer, researchers have been able to vary the properties of speechlike sounds and so investigate the phonetic cues that listeners use. It turns out that listeners often do not depend on a single cue but rather weigh the evidence from several independent features. Some results have been particularly surprising. For example, the words *ones* and *once* are normally felt to differ just in their last phoneme, *ones* ending in the voiced phoneme /z/ and *once* in the corresponding voiceless phoneme /s/ (in *voiced* sounds the vocal cords act as a quasi-periodic sound source; in *voiceless* sounds they do not); but it is possible to change a listener's judgment of which word he is hearing merely by altering the length of the /n/ sound (a longer /n/ indicating *ones),* and indeed it seems likely that this is the most important phonetic cue in discriminating between these words in natural speech. Here we have an example, then, where the major distinguishing mark of a phoneme is not only not what we would expect it to be, it is not even *where* we would expect to find it.

Moreover, some work carried out in England [1] has shown that cues to phoneme identity are not even entirely confined to the auditory channel: in appropriate circumstances visual cues can be integrated into speech perception. The point has been convincingly demonstrated by synchronizing a recording of a stop consonant vowel sequence e.g. "ba" with a video recording of a person producing a different stop consonant followed by the same vowel - e.g. "ga". The perception of the sound is strongly modified by the conflicting visual cues - in the ba/ga example what is perceived is "da". The effect has perhaps to be seen to be fully believed: when I saw the demonstration I "heard" a perfectly natural "da" whilever I watched the screen; it reverted to "ba" as soon as I heard it while looking away from the screen.

I hope all this is beginning to convince you that speech cannot be considered as a sequence of speech sounds in the way that the teleprinter transmission is a sequence of teleprinter symbols.

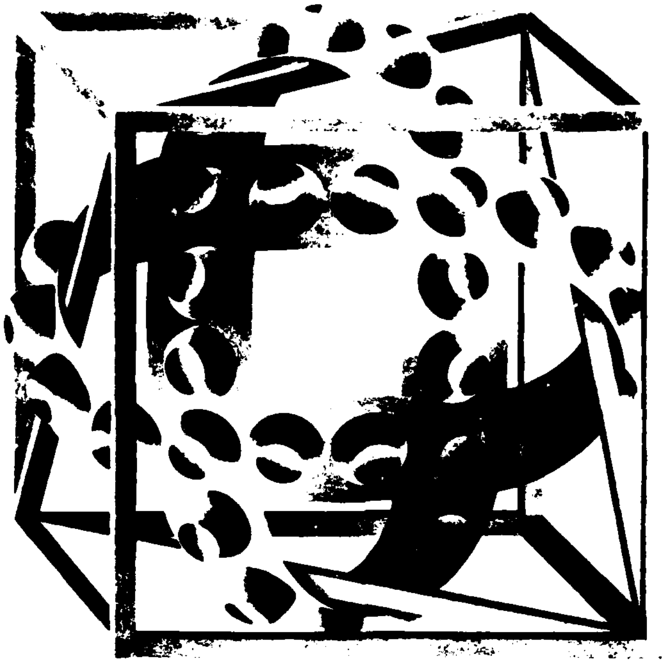


**Fig. 1** A visual paradox: *Cube with Magic Ribbons* by M.C. Escher (courtesy M.C. Escher Foundation, the Hague).

The active nature of speech perception

Before I go on, I would like you to look at the M.C. Escher drawing reproduced in Figure 1. It seems paradoxical: among other problems that it poses, the circular objects on the ribbon seem to flip from pointing outwards to pointing inwards. If we were able to regard it as a meaningless pattern of different shades of grey on a flat piece of paper, there would be no paradox. But it seems all but impossible to restrain our minds from attempting to reconstruct a three-dimensional object out of the pattern, even when such a reconstruction cannot be made to work. The picture illustrates the point that visual perception does not work simply by recording the light entering the eye, but rather by actively trying to "make sense" of that light. To give another example, we can perceive the color brown, but there is no such thing as brown light: apparently our brain deduces the "brownness" of an object by comparing the quality of the light reflected from the object to that of the light that our brain computes to be striking the object.

I want to suggest that our hearing is similar: it no more works like a microphone than our vision works like a camera. In particular, listening to speech is not a passive detection of an acoustic signal: it is an *active reconstruction* of the transmitted message.

This reconstruction is so effective that we frequently do not notice that apparently important information is missing: until we have to deal with something unfamiliar like an unusual name, we are hardly aware that it is not possible to distinguish between an "s" sound and an "f" sound on the telephone; and synthetic speech in which all the voiceless sounds have been replaced by silence seems surprisingly normal, particularly if there is some background hiss that our brain can take to be voiceless fricatives.

It is the reconstruction that gives us such a firm impression that the speech signal consists of a neat sequence of phonemes: it may indeed be possible to describe speech in this way, but only at a certain stage of processing in our brains, not at the level of the acoustic signal.

In reconstructing the speech message the listener can use information from several different sources. We have already mentioned prosodic cues such as intonation that generally indicate sentence structure, and phonetic cues that indicate word structure. There are rules that govern the order in which words can be uttered in the *syntax* of a language, and other constraints labeled as *semantics* that provide that most sentences should be meaningful. Syntax and semantics are clearly linked, but they are at least partially separable: Chomsky's sentence *colorless green ideas sleep furiously* is syntactically acceptable but meaningless, while *Me Tarzan, You Jane* breaks the rules of standard English syntax but was nevertheless meaningful to the cinema audiences that heard it. To our list of information sources we might also add *external context*, that is, whether an utterance is germane to the situation, and whether it is the sort of remark the speaker frequently makes under his present circumstances. Finally, the work with synchronized video recordings demonstrates that in some circumstances optical information is used in reconstructing the speech message.

This leads me to point out another way in which speech differs from the teleprinter transmission, namely, the fact that speech has to be regarded as a *multilevel* sequence. Thus, words can be thought of as phoneme sequences, while they themselves form part of word sequences making up phrases, which in turn make up sentences. Evidence needed to understand speech is present at every level, and in all probability the evidence at all levels has to be considered simultaneously if the message is to be understood. It is true that we could find much the same set of levels in a teleprinter transmission of meaningful text, but the levels are not so intimately mixed: in order to decode the individual teleprinter symbols we do not even need to know what language the text is written in.

It is often said that speech is a very redundant signal. As evidence for this assertion, it might be pointed out that the same utterance can be understood either when it is low-pass filtered at 1kHz or when it is high-pass filtered at 1kHz: the information in the lower part of the spectrum must somehow be duplicating the information in the upper part. I believe this to be a fallacious way of looking at the speech signal. The amount of information one needs in a speech signal depends on how skilled one is at reconstructing the message: I need much higher signal quality to follow spoken French or German than I do to follow spoken English. To mangle a metaphor: redundancy is in the ear of the beholder.

This brings me to what is perhaps the most important point in this talk. It is that people do not emit speech messages to be picked up by anyone who cares to listen, they *talk to someone*. Although we as yet know too little about speech to be sure about this, it seems likely that a speaker puts just enough cues into his speech to allow his listener (or imagined listener in the case of, say, a radio broadcast) to be able to comfortably reconstruct the message from the evidence available. Thus, when we are saying something that is difficult to follow, or when we are speaking to someone we believe to be foreign, deaf or senile, we supply more phonetic information than we would in a relaxed conversation with a friend. Elision of phonetic information, such as when we say *fish 'n chips*, is often described as being due to laziness, but I would argue that it is part of a rational strategy for the economical use of a communications link: it would be lazy only if the person at the other end of the link were obliged to make an unreasonable effort to reconstruct the message. Depending on the circumstances, overarticulation can be just as inappropriate as underarticulation: it can sound stilted, irritating, even insulting when the listener feels it to be unnecessary.

To summarize my account of the speech signal so far, I have tried to argue that it is different in nature both from the messageless signals we considered first and from machine-generated message-bearing signals like the teleprinter transmission. It is a signal from which a message may be reconstructed using information drawn from many sources, both information at various levels in the signal itself and information stored in the mind of the listener. The amount of information that the speaker puts into the signal depends on the difficulty that he imagines the listener will have in reconstructing the message from it.

### The speech signal and speech recognition

I would like to turn now to considering the speech signal more specifically in relation to automatic speech recognition.

#### The use of syntactic constraints

Because we are directly aware of speech only after it has been subjected to extremely sophisticated processing involving information from several sources, it is all too easy to underestimate the difficulty of deducing a spoken message purely from the acoustic signal. In particular, there is a danger of expecting to find products of a high level analysis of speech such as phonemes to be present as clearly identifiable entities in the acoustic signal. Presumably, it was false impressions such as these that influenced the author of a recent market study of speech technology [2] when he predicted that commercially viable voice activated typewriters would be available in 1987, the limiting factor, according to him, being the cost of memory to store a large vocabulary.

It is not clear to me that it would ever be possible for a machine to recognize unrestricted speech with high reliability purely from the acoustic signal - it is, at best, like asking a human listener to transcribe accurately a language he does not understand. Accurate transcription of unrestricted text probably requires both a knowledge of the syntax of the language and a comprehensive knowledge of the world. Present day practical systems limit the difficulty they face either by having a small vocabulary or by having a larger vocabulary but with a syntax that limits the choice of words that can follow a previous word or sequence of words. For it is, of course, the number of *choices* that the system has to discriminate between that determines the difficulty of a recognition task not the total number of words it has to recognize. To give a trivial example, the task of recognizing the two word vocabulary *Paris* and *London* is made easier if we go to a four word vocabulary by including the words *France* and *England* together with a syntax that requires that the city must be followed by its corresponding country.

Some of the most ambitious systems using syntactic and semantic constraints were constructed as part of the ARPA Speech Understanding Project [3]. Systems were devised that had "knowledge" of a small subset of the syntactic structures possible in English and an "understanding" of a small universe (such as facts about ships). Since that project ended in 1976, however, there seems to have been a considerable reduction in interest in such systems. As I see it, apart from the high cost, there are several good reasons for this loss of interest. First, there is the problem of the very considerable effort needed to specify the syntax and semantics of the language to be used for a particular application. This prevents speech understanding devices from being sold as off-the-shelf devices. Second, there is a problem in defining what is known technically as a *habitable subset* of a natural language. That is to say, as the syntactic structures allowed by a system get more complex and the language one can use gets more natural, it gets correspondingly harder to teach a user what sentence structures are grammatical to the recognizer as opposed to those that are grammatical in the user's own language but not allowed in the recognizer's grammar. Finally, as a research tool complex systems seem to me to be unattractive because when overall performance depends on so many factors it is difficult to draw useful conclusions from that performance or from the relative performance of two such systems.

I wonder if there is perhaps a parallel to be drawn between speech recognition devices and robots. Before any useful robots had been built the image of the robot was of a device that superficially resembled a man; but real, useful robots working for example in car factories do not look at all like humans. Real, useful speech recognizers do not use a syntax that superficially resembles natural language, though they do increasingly use a task-oriented syntax.

An example of a very simple yet effective use of task-oriented syntax in a recognizer is the addition of a check digit to a string of digits to be recognized. This digit would typically be chosen such that when a string of digits including the check digit is summed together the result is always a multiple of ten. (For instance, the string 1 1 1 would have 7 as a check digit, while 8 8 0 would have 4.) The inclusion of the check digit does not reduce the total number of possible digit strings that the system has to discriminate between - if we have a three-digit string there are a thousand possibilities whether we add a fourth check digit or not - but it does increase the amount of acoustic information that can be used in the discrimination. Alternatively, we can view the check digit as having made discrimination simpler by reducing the average number of choices to be made per word. This average number of choices is known as the *branching factor*, and it - or a generalization of it when the choices are not equiprobable - is often used as a measure of difficulty of recognition tasks.

The use of devices like check digits is not as alien to natural language as it might appear. A similar recognition-aiding device occurs, I believe, in all Indo-European languages except the one I am using now. I am referring to the division of nouns into two or three classes according to what is called the *gender* of the noun, the gender classes being called variously masculine, feminine and (sometimes) neuter, or neuter and common. To illustrate how it can help, consider the French nouns *poisson* (fish) and *boisson* (drink) that are quite similar in pronunciation, but differ in that the first is masculine and the second feminine. When we meet them in sentences like

*Le X est un poisson délicieux* and *Le X est une boisson délicieuse*

(X is a delicious fish/drink) it is virtually impossible to confuse them despite their phonetic similarity because the form of the adjective and the indefinite article both depend on the gender of the noun they refer to and are therefore different for *boisson* and *poisson*. In French there are only two noun classes against ten check digits, so instead of calling gender a check digit we should perhaps better describe it as a linguistic parity bit, but the principle is the same.

## Approaches to speech analysis

So far, I have tried to point out some of the difficulties in analysing the speech signal and the dangers of methods based on introspection. I would like to look now at some approaches that have proved helpful. Useful approaches to the treatment of the speech signal seem to fall under three headings, namely, production based approaches, perception-based approaches and pragmatic approaches. No automatic recognition system relies totally on just one of these approaches, but in most systems one approach dominates.

Production-based approaches

It does not seem immediately obvious why we should approach the *recognition* of speech from the viewpoint of how it was produced - we do not, after all, need to know how the teleprinter signal was generated in order to decode it. Nevertheless, there is a whole spectrum of arguments in favor of taking speech production into account when analysing speech. They range from the most moderate, with which no one would argue, to the most extreme, which few people now hold.

Before we consider these arguments, I shall have to break off from my main line of argument for a little while in order give you a brief overview of how speech is produced. The human organs primarily involved in producing speech are the *larynx,* which contains the vocal cords, and the pharynx and mouth cavity, which together form the *vocal tract,* and which is essentially a tube leading from the larynx to the lips. A side branch, the *nasal cavity,* can be added to this tube by opening a valve at the back of the mouth. This valve is open in nasal consonants, such as an "m" sound, and in nasalized vowels, which form a separate class of phonemes in some languages such as French.

Acoustic energy in speech is generated in one of two ways: by the action of the vocal cords or by turbulence at a constriction created by the tongue or lips somewhere along the vocal tract. As I mentioned earlier, sounds excited by the quasi-periodic activity of the vocal cords are said to be *voiced,* and they generally play a more important role in speech than noise-excited *voiceless* sounds. (All vowels and many consonants, such as "l", "m" and "b" sounds are voiced, while "sh", "k" and "f" are examples of voiceless sounds.)

Whichever kind of excitation is used, the basic spectrum of the excitation is modified by the resonant structure of the vocal tract. This resonant structure depends on the position that the tongue, lips and jaw are in. It happens that the generation of the excitation and its spectral modification by the vocal tract are largely independent of each other and can thus be considered to a good approximation as a source isolated from, and leading into, a linear filter.

The upper trace of Figure 2 shows a 200ms stretch of the waveform of a non-nasalized vowel (strictly, it is the *time-differenced* waveform: differentiation provides a 6db per octave lift, which serves to flatten the long-term spectrum for voiced speech). Notice that the waveform consists of a pattern that repeats itself at regular intervals. The repetition rate is the rate at which the vocal cords come together - typically a hundred times a second for a man - while the repeating pattern itself is the response of the vocal tract to that periodic excitation.

The lower trace in Figure 2 shows the excitation with the effect of the vocal tract removed. The impulse-like excitation occurs each time the vocal cords come together and close off the airflow from the lungs. In the particularly simple vowel shown here (it is in fact the "neutral" vowel occurring in a word like "the") essentially what happens to the impulse is that it travels from the larynx to the lips, where part of it is radiated into the free air beyond the lips and part is reflected back towards the larynx with its polarity reversed. At the larynx the signal is reflected again, and it continues to bounce between larynx and lips steadily losing energy by absorption in the walls of the vocal tract, by transmission and ultimate absorption behind the vocal cords, and by radiation to the outside world until the next excitation impulse comes along. In other speech sounds the effect of the vocal tract on the excitation is more complex, with reflections occurring at more places than just the larynx and lips. Nevertheless, the basic structure of a pattern approximately repeating itself at approximately regular intervals is retained.

Figure 3 shows the power spectrum of a section of speech waveform like the one in Figure 2. The regularly spaced spikes occur at integer multiples of the repeat frequency of the excitation. This repeat frequency, corresponding to the first spike in the figure, is known as the *fundamental frequency,* and the succeeding spikes are *harmonics* of the fundamental. The intensity of the harmonics varies smoothly across the spectrum in a way determined by the *impulse response* of the vocal tract. The peaks in the spectrum coincide with resonances in the vocal tract. They are known as *formants.*

The ability to describe the speech signal in terms of an impulse response and the frequency of the impulses is extremely important for speech recognition. The impulse response varies as the positions of the tongue, jaw and lips are changed, while the fundamental frequency depends on the muscles that control the tension in the vocal cords and on the air pressure behind the vocal cords. For the most part, changes in the settings of the larynx and vocal tract occur slowly relative to the frequencies involved in speech. Thus, while we need a sampling rate of at least eight thousand

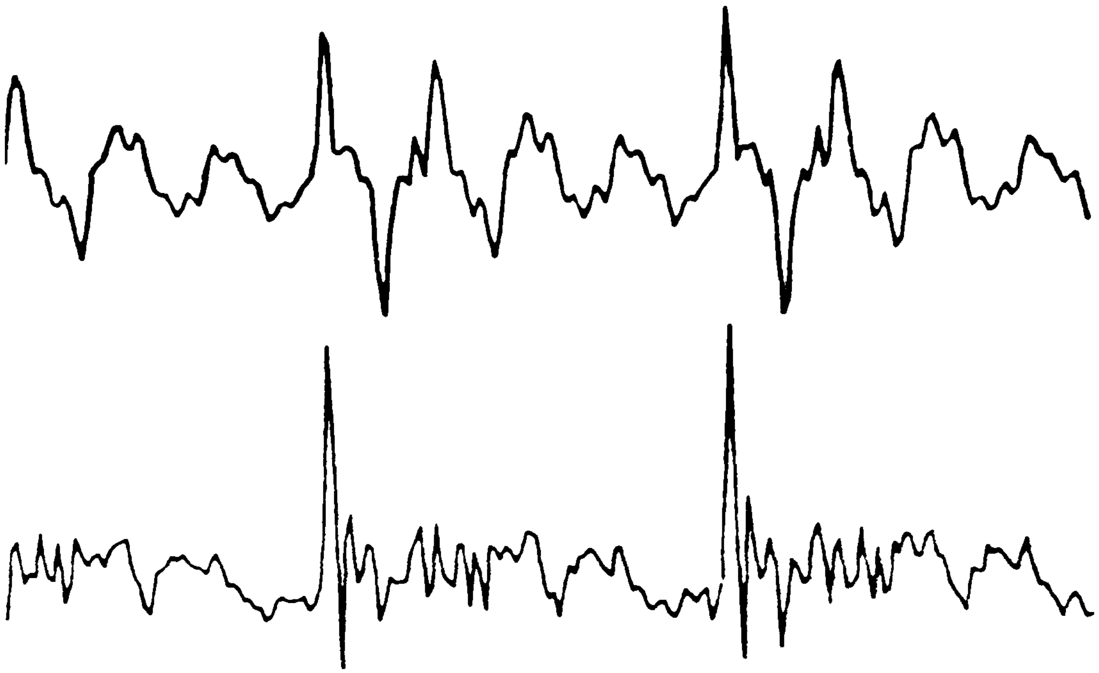


**Fig. 2.** Upper trace: a 200ms portion of the time-differenced waveform of a neutral vowel. Lower trace: the same waveform with the effect of the vocal tract removed.

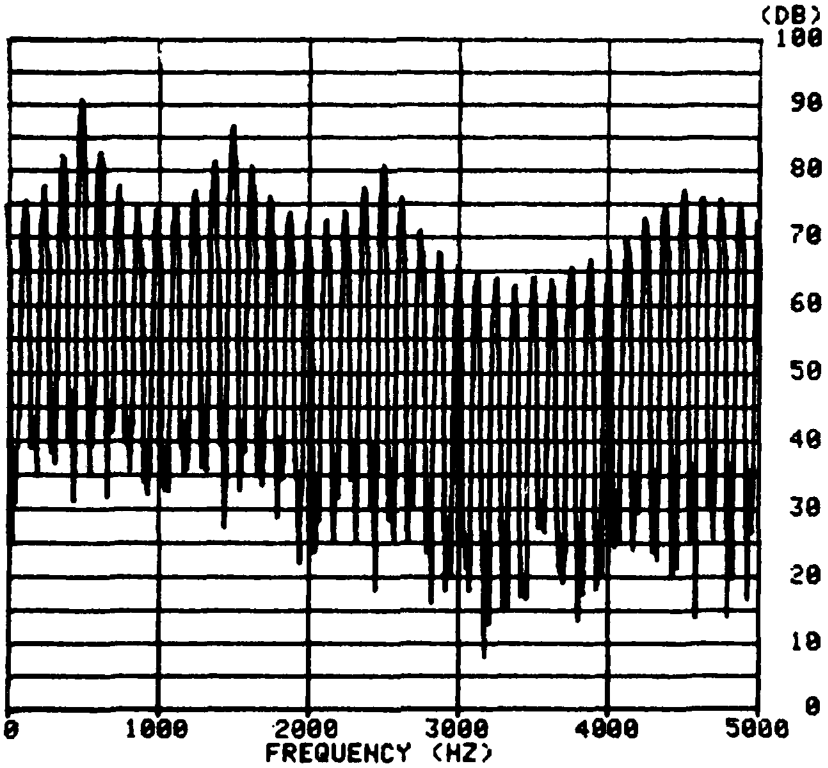


**Fig. 3.** The power spectrum of a time-differenced neutral vowel.

times a second in order to obtain a reasonable digital description of the speech waveform, a description in terms of fundamental frequency and a few parameters describing the impulse response typically needs to be updated as little as a hundred or even fifty times a second.

A second major advantage of an impulse-response/fundamental frequency description is that the two factors perform separate linguistic functions. In most western languages the identity of a word does not depend on the fundamental frequency pattern with which it is uttered. In some other languages, such as Chinese and to a much lesser extent Norwegian, the identity of a word may depend on the fundamental frequency pattern, but even then a practical recognition strategy must still separate the two factors: the fundamental frequency pattern and the configuration of the articulators in the vocal tract are substantially independent attributes of the word.

For non-nasalized vowels and some non-nasal consonants the impulse response of the vocal tract is quite accurately modeled by a set of resonances in series, the important resonances lying in the range 300Hz to 3kHz. For such sounds, provided the analysis is carried out in proper synchrony with the excitation, a technique known as *linear prediction* can be used to determine from the waveform the frequencies and bandwidths of the resonances (a comprehensive account of linear prediction is given in the book by Markel and Gray [4]). What is more, the analysis can go on to reconstruct the cross-sectional profile of an acoustic tube that would have such a set of resonances, and the profiles can often show a close similarity to the vocal tract configuration that produced the sound. This kind of analysis of speech can therefore be said to be strongly production-oriented. I should point out that that the analysis is approximate in as much as the model of the excitation by a sequence of impulses is inexact, that the analysis is inevitably less successful for certain sounds such as nasals where the simple-tube vocal tract model does not fit, and that in most practical cases its accuracy is further reduced by applying the analysis at regular intervals along the waveform rather than in synchrony with the excitation.

We can now get back to considering the arguments in favor of approaching speech recognition from the point of view of speech production. We can see that at the moderate end, proponents of production-based analysis could point out that it can lead to the generation of a simple, compact description of the speech signal, and, moreover, one in which features that determine the lexical identity of a word are quite well separated from features that have more to do with the function of the word in the sentence or with mood of the speaker. Somewhat more speculatively, if we could carry out a production-based analysis well enough we might hope to predict coarticulation phenomena such as our *breab board* example as well as other energy-saving shortcuts that the articulators might take, such as the possible failure of the tongue to reach its target position for a vowel separating two consonants in a rapidly spoken syllable. Finally, the most extreme view, embodied in the *Motor Theory of Speech Perception* [5], maintains that human speech perception works by mentally reconstructing the articulator settings that produced the speech signal being heard. According to this last view, which is, I believe, much less popular than it was fifteen years ago, it would be highly desirable that an automatic recognizer should also work by reconstructing the production details of the speech signal it is receiving. Linear prediction looks to be the best current possibility for carrying out such a reconstruction.

### Perception-based approaches

To some people it may seem a truism to assert that we should base speech recognizers on human speech perception. Others sometimes point out that aircraft don't flap their wings just because birds do, so machine recognition of speech need not copy the human model. But this is a false analogy: first there was air, and then birds and men found ways of flying in it; I doubt if anyone would claim that first there was speech and then men evolved ears to listen to it! Speech is certainly adapted to our human capacity to perceive it, and it is desirable that an automatic recognizer should be able to make the distinctions that we can make and ignore those we cannot make.

The problem in basing a recognition approach on human speech perception lies not in deciding whether it is a good idea to do so, but rather in the fact that we know remarkably few hard facts about human speech perception. What we know somewhat more about is human *sound* perception in general. We know, for example, that we are relatively insensitive to the phase information in a signal, and it would consequently make little sense to build a recognizer that tried to recognize a specific waveform, since a slight change in the relative phases of its spectral components would be indetectible to human ears and yet it could cause the waveform to look quite different. For this reason, all practical recognizers work with some representation or other of the short-term power spectrum and ignore the phase spectrum.

Another known property of human sound perception is *frequency masking*, the tendency of an intense tone to obscure the presence of a less intense tone at a neighboring frequency. It follows from this property that our hearing is more sensitive to the peaks in the spectrum than to the troughs, whose details tend to be masked by nearby peaks. Thus a speech recognizer that used a representation of the spectrum that was particularly sensitive to dips in the spectrum would be unlikely to work well. If the algorithm commonly used in linear prediction is viewed as a means of characterizing the short-term power spectrum, it turns out that it has the desirable property of doing a better job of characterizing the peaks than the troughs.

An alternative - and in fact much older - method of reducing sensitivity to weak tones when they are close to a strong tone is to divide the auditory spectrum into a set of bands, the acoustic power in the set of frequencies in each band being averaged together. Masking experiments with human subjects tell us how wide these *critical bands* should be: above about 1kHz the bands should not be of constant width, but rather they should increase in width in rough proportion to their centre frequency with about three bands in each octave. A *channel vocoder* uses a bank of filters that is rather like the set of critical bands, and the same principle is used in many successful speech recognizers. There is an interesting conflict between production-based and perception-based approaches here. A representation of speech based on linear prediction, which generally models speech production quite well, gives equal resolution to all parts of the. spectrum. Thus, a comparison between two recognition systems which were similar except that one carried out a filter-bank analysis of the speech and the other a linear-predictive analysis would amount to a comparison between a production-modeling approach and an auditory-perception modeling approach. Davis and Mermelstein [6] carried out such a comparison and found a clear advantage for the filter bank. Moreover, among the parameter sets that can be used to present the results of the linear prediction the ones that are best interpreted as providing a description of the general shape of the spectrum (i.e. the linear prediction *cepstrum)* performed better than the more production-oriented area coefficients that would be used in reconstructing vocal-tract area functions.

Some researchers [7,8] have gone further in modeling neural behavior in the inner ear, in some cases incorporating the superior time resolution available to the ear at high frequencies where frequency resolution is poor. Although improved recognition of stops and fricatives has been claimed to result, the procedure has not been widely adopted because it is computationally expensive.

If we now turn our attention to *speech* perception rather than auditory perception, I have to admit that I find the field confusing, and I do not feel competent to make any attempt at an overview of present knowledge. There is, perhaps, evidence [9] that speech processing works in a "left-to-right" fashion (i.e. forwards in time) rather than, say, first picking out stressed syllables and working outwards from them in both directions, and that possibilities for each word to be recognized are considered in parallel rather than exploring the most promising interpretation first and returning when it meets trouble. I am sure, however, that both these statements would be disputed by some specialists in speech perception.

In 1979 Klatt published a long paper [10] proposing the incorporation of models of human speech perception in a recognition system. He subsequently reported experiments [11] suggesting that listeners use a different criterion when making a judgment of the *phonetic* similarity of two sounds from the one they use when making a general psychophysical comparison of two sounds. The psychophysical judgments seem consistent with the general spectral shape comparisons carried out in most recognition systems, while the phonetic judgments seem more dependent on the frequencies of energy peaks in the spectrum. He has more recently reported work on a metric intended to correlate better with human phonetic judgments [12]. It will be interesting to see how the metric performs - many researchers in the past have thought it desirable to represent speech in terms of the frequencies of energy peaks - formant frequencies - but have been held back by the problem that occasional errors in peak frequency assignment can have disastrous results on performance. The new metric avoids making hard decisions about formant frequencies.

In general, it is striking how little the results of research in speech perception have influenced the design of successful speech recognition systems, though that does not, of course, preclude such influence in the future.

### Pragmatic approaches

The heading "pragmatic approaches" seems at first sight like a catch-all under which any recognition work not based on production or perception results can be placed. To some extent it is just that, except that it excludes approaches justified by introspection or by pet theories inadequately supported by experimental evidence. I mean the term to be confined to approaches that are justified primarily by the fact that they are found to *work*. A notable example of such an approach is provided by the various versions of the *dynamic programming* algorithm for time-aligning two productions of a word or sequence of words. The algorithm will no doubt be explained in detail in later sessions; for the moment all I want to point out is that it is central to a large proportion of successful recognition systems and that its introduction was not inspired by production or perception considerations but rather by the fact that it could cope with a phenomenon found to occur in the signal, namely, non-linear timing variations amongst different productions of the same word.

Perhaps the most extreme example of an approach based directly on the properties of the signal itself is the work of Jelinek's group at IBM [13]. Instead of having syntactic rules supplied by the system designer, the system itself deduces transition probabilities between words from a very large amount of training data. It then uses those probability estimates in attempting to decode new material. Results have been reported on a database of natural English consisting of a set of patent texts concerning lasers. It constitutes the most ambitious current attempt at single-speaker speech recognition that I know of.

One property that many of the more successful - and above all, practically useful - systems share is *simplicity*. I suspect it is no accident that *Harpy*, the only one of the ARPA Speech Understanding systems to meet the original success criteria, was distinguished from its competitors mainly by the fact that it was considerably simpler. John Bridle's successful continuous word matching algorithm [14], which I hope he will describe to you, is also considerably simpler than other algorithms that have been proposed for recognizing word sequences.

Why should simple approaches be better? I think the main reason is that they have fewer system parameters to tune, and they can consequently reach a better state of optimization with a given amount of training data than could a more complicated system. A second, related, reason is that in developing a simple system a developer can more easily assess the effect on performance of a design decision he has made than he could in a complicated system in which many rules and processes interact.

I am convinced that for the foreseeable future practically useful recognition systems will remain simple systems. To the extent that their design reflects properties of human speech production or perception, I believe that the better ones will be based on solidly established properties and not on speculation.

## Further reading

If anyone is interested in a closer look at phonetics, there are introductory texts by Ladefoged [15] and O'Connor [16]. The standard work on the acoustic theory of speech production was written by Fant [17].

## References

[1] McGURK H. & MacDONALD J. "Hearing lips and seeing voices", *Nature* Vol. 264 #5588, pp.746-748, 1976.

[2] WEISSMAN I., *IRD Market Survey No. 516: the U.S. Speech Recognition and Synthesis Market* as reported in *Voice News* Vol. 3 #1 Jan. 1983.

[3] KLATT D.H. "Review of the ARPA speech understanding project", *J. Acoust. Soc. America*, Vol. 62, pp. 1345-1366, 1977.

[4] MARKEL J.D.& GRAY A.H. *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976.

[5] LIBERMAN A.M, COOPER F.S, HARRIS K.S. & MACNEILAGE P.F "A motor theory of speech perception", *Proc. Stockholm Speech Comm. Seminar*, R.I.T., Stockholm, September 1962.

[6] DAVIS S. & MERMELSTEIN P. "Evaluation of acoustic parameters for monosyllabic word recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-23, pp.82-87, 1975.

[7] SEARLE C.L., JACOBSON J.Z & RAYMENT S.G. "Stop consonant discrimination based on human audition", *J. Acoust. Soc. America*, Vol. 65 pp.799-809, 1979.

[8] ZWICKER E., TERHARDT E. & PAULUS E., "Automatic speech recognition using psychoacoustic models", *J. Acoust. Soc. America*, Vol. 65, pp.487-498, 1979.

[9] MARSLEN-WILSON W.D. "Speech understanding as a psychological process", in *Spoken Language Generation and Understanding*: Proceedings of the NATO Advanced Study Institute held at Bonas, France, June 26-July 7, 1979, D. Reidel Publishing Co., Dordrecht, 1980.

[10] KLATT D.H. "Speech perception: a model of acoustic phonetic analysis and lexical access", *Journal of Phonetics*, Vol. 7, pp.279-312, 1979.

[11]KLATT D.H., "Perceptual comparisons among a set of vowels similar to /ae/: some differences between psychophysical and phonetic distances" *J. Acoust. Soc. America*, Vol 66 p.S86, 1979

[12]KLAT D.H., "Prediction of perceived phonetic distance from critical-band spectra : a first step" *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, May 1982, pp.1278-1281.

[13]BAHL L.R., BAKIS R., COHEN P.S. COLE A.G., JELINEK F., LEWIS B.L. & MERCER R.L. "Further results in the recognition of a continuously read natural corpus", *P roc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Denver CO, pp.872-875, April 1980.

[14]BRIDLE J.S., BROWN M.D. & CHAMBERLAIN R.M. "An algorithm for connected word recognition", *P roc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, pp.899-902, May, 1982.

[15]LADEFOGED P. *A Course in Phonetics*, Harcourt Brace Jovanovitch Inc., New York, 1975.

[16]O'CONNOR J.D. *Phonetics*, Penguin Books, London, 1973.

[17]FANT G. *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations*, Mouton, the Hague, 1970.

# ANALYSIS, SYNTHESIS AND TRANSMISSION OF SPEECH SIGNALS

Helmut Mangold
AEG-TELEFUNKEN, Research Institute Ulm
Sedanstr. 10
D-7900 Ulm (West Germany)

## SUMMARY

Digital techniques have opened quite new possibilities for processing of speech signals. This is true for analysis and for transmission. These new methods are characterized by a strict adaptation to the very special pecularities of speech.

The lecture will give an overview about the mathematical possibilities and their relevance to the different parts of the speech signal. Efforts to represent speech in a digital and more or less redundancy-free form can give good insight into all the characteristics of such a highly complex signal.

Possibilities for representation of speech signals reach from the very simple pulse-code-modulation techniques (PCM) to sophisticated vocoders.

The research work done for speech transmission and coding has prepared the way for methods to recognize and synthesize speech signals. Automatic speech synthesis will be an important tool for the communication between man and machine. The lecture will give an additional introduction into the techniques of automatic speech synthesis.

## 1. INTRODUCTION

Speech signals are the most important signals in today's and tomorrow's telecommunication systems. This results from the fact that human communication is the basis of all communication systems. This man-to-man communication will in the future be combined with efficient man-machine communication. About one part of this communication, speech output by computers, later in this paper will be reported.

Techniques of digital signal processing have opened quite new and exciting ideas, how to handle the structure of speech signals. We can now describe quite well the information-theoretic content of the signal. Most of the characteristics which are necessary to develop a suitable model for the speech signal can be understood by the principles of natural speech production /1/. Fig. 1 shows the essential parts of this process. In the case of voiced sounds a pulse exitation signal is produced by the vocal cords which are vibrating when the air stream from the trachea passes through. These pulses are modulated within the cavities of the throat, the mouth and the nose and the resulting signal will be a periodic voiced signal. Its characteristics, that means the sort of sound is defined by the acoustic porperties of these filtering cavities. Their properties are defined by the geometric dimensions of the cavities which can be changed during the process of articulation.

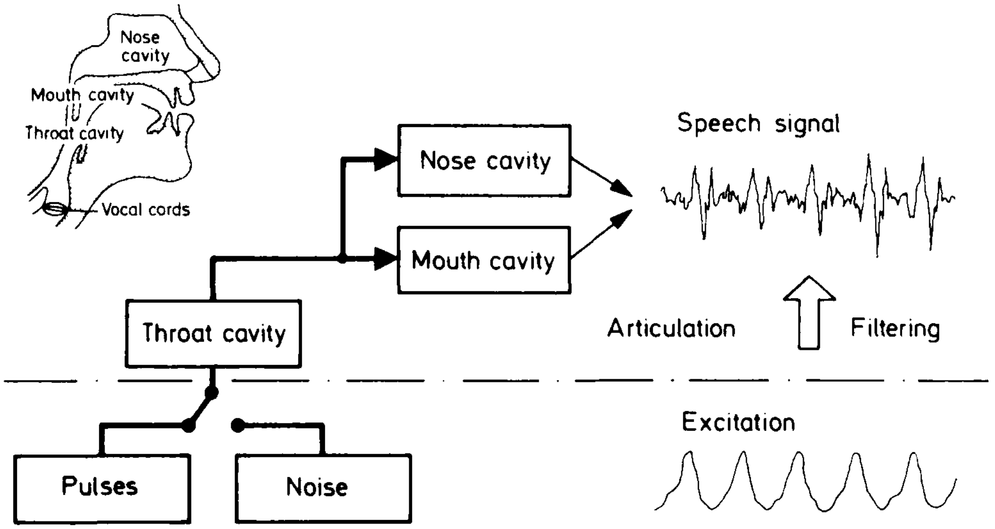


Fig. 1: Principles of natural speech production.

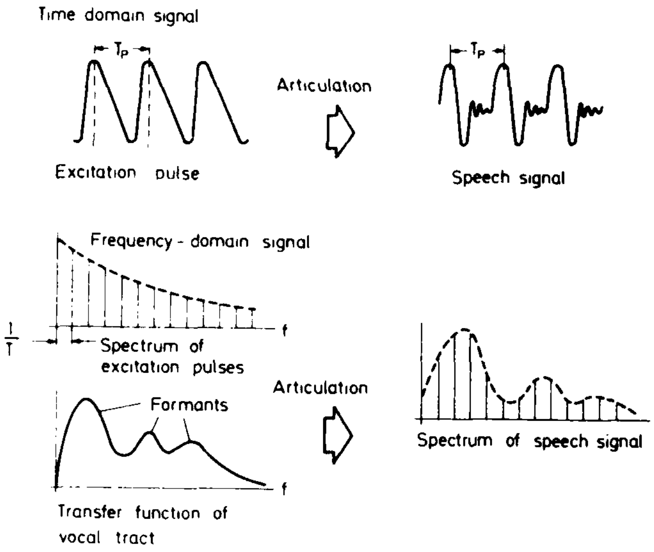


Fig. 2: Time-domain and frequency-domain representation of speech.

Fig. 2 gives an overview about the important signals in the process of speech production and speech perception. Fig. 2a shows the already described time-domain signal. The pitch of the excitation pulses $T_p$ is about 120 Hz in the average for male voices. We can find this period again in the scheme of the speech signal which is characterized by higher frequency waveforms, the so-called formant frequencies. This might be understood much better if we look at the frequency characteristics of such a speech signal (Fig. 2b). The excitation pulses of voiced signals have a line spectrum with an envelope that falls to higher frequencies with about 6 to 10 dB/octave. The spectral transfer function of the vocal tract has very strong resonances, the already mentioned formants, which characterize the sound of the speech signal. During the process of articulation the spectral envelope of the excitation signal is modulated by the transfer characteristics of the vocal tract, which has sharp resonance peaks. For unvoiced sounds the basic process is quite similar. The only difference comes from the fact that the excitation signal now consists of a sort of turbulence noise which is created by air, streaming through some quite narrow positions within the vocal tract.

The process of speech production is of course a dynamic process. That means that all the mentioned parameters are changed in a relatively fast manner. The normal speaking rate is about 10 to 20 sounds per second. The duration of different sounds varies between 5 ms for the very short plosive sounds like /t/ up to about 100 ms for slowly spoken voiced sounds like some vowels. These timecharacteristics of speech signals are important for many speech analysis and synthesis techniques. In the case of speech transmission it is additionally important to know something about the human perception of speech signals.

Man's acoustic perception system is not exclusively dedicated to the perception of speech. However its perception principles are very well adapted to the special qualities of speech signals. In principle the ear makes a spectral analysis with additional emphasis on the analysis of time-varying signals. This means that there is a combination of a sort of very narrow band spectral analysis for precise detection of the formant's mid frequencies and simultaneously a precise analysis of time variations in the spectral characteristics including periodicity detection for the analysis of the varying line structure of voiced speech signals. So our speech percep tion appartus is a highly sophisticated system with special adaptation to the structure of speech signals. We must take care of these and many more facts if we want to design good speech transmission systems and want to produce natural and intelligible speech. On the other side speech signals have been optimally adapted to a sort of spectral anlysis with quite special poperties which is done within our human ear and the following neural stages in our brain. So technical systems for speech analysis not only must take care of the physiological processes but also can learn many things from these processes. These ideas are especially important for preprocessing and feature extraction stages in a speech processing system.

Fig. 3 gives us a rough overview about this interaction of speech transmission and recognition/synthesis ideas. Every technical analysis starts with a sort of preprocessing by which some important signal characteristics are extracted. The following feature extraction is still a preprocessing stage for a speech recognition system but there are extracted more complicated and combinatorial parameters, e.g. segmented phoneme parameters or prosodic parameters like speech intonation. The following stages are concerned with the central task of recognition and understanding. Then a speech output is created based on linguistic rules. The phonetic and speech synthesis parts again handle higher and lower level parameters to produce a speech signal which will be put to a loudspeaker

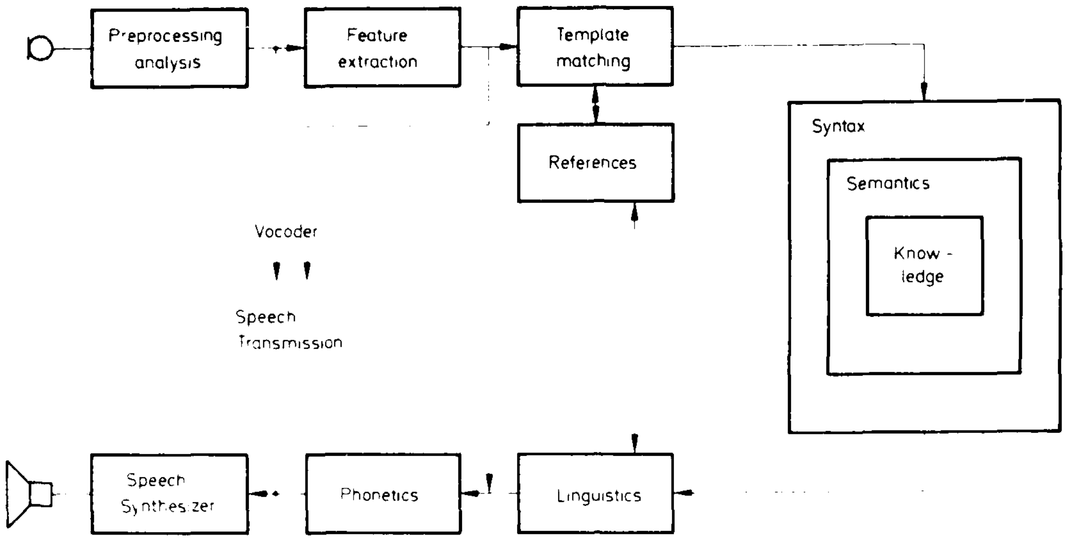


Fig. 3: Relations between speech transmission and speech recognition and synthesis.

to make an acoustic signal. In speech transmission with redundancy reduction we jump over the inner kernel of this system from Fig. 3 and transmit directly a parametric description of the analyzed signal to a sort of synthesizer which can reproduce the physical signal.

So it will be important to understand the significance of different sorts of preprocessing in speech processing by studying some of the more important speech transmission systems today in use and to understand how their signals can be useful for automatic speech recognition and synthesis.

## 2. MATHEMATICAL AND THEORETICAL PRINCIPLES OF DIGITAL SPEECH PROCESSING /4/,/5/

The term "speech processing" does not automatically include the term "digital" but in practice today analog speech processing is still only used in very special cases. So the basis of all our operations will be a sampled and quantized signal. This means that the speech signal has to be coded into a form of numbers. The principle of this pulse code modulation (PCM) process is shown in Fig. 4 /2/. The analog waveform (Fig. 4a) is sam-

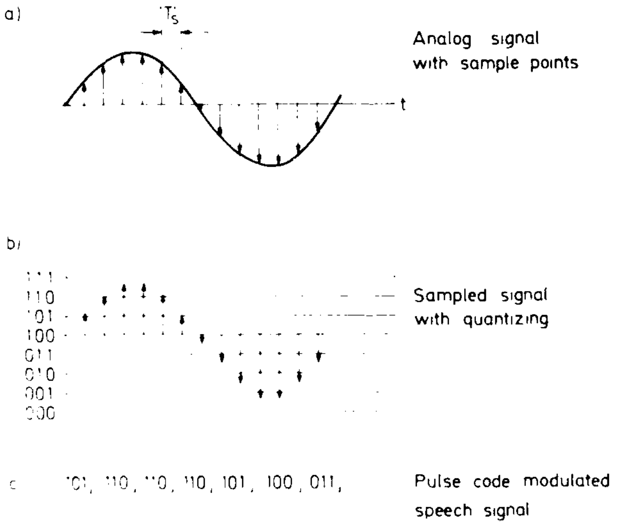

a)
$T_s$
t
Analog signal with sample points
b)
111
110
101
100
011
010
001
000
Sampled signal with quantizing
c
101, 110, 110, 110, 101, 100, 011,
Pulse code modulated speech signal

Fig. 4: Puls-code representation of analog signals.

pled at a fixed rate $1/T_S$ /Hz/, normally with double the highest frequency which is in the signal. Telephone quality speech has a bandwidth of about 4 kHz, so it is necessary to sample such a signal with at least 8 kHz. This means that $T_s$ is 125 μs. Speech signals with better quality need a higher sampling rate up to 20 kHz, resulting in a speech bandwidth of 10 kHz. Sampling produces an amplitude modulated impulse signal (Fig. 4b). The amplitude of every of these impulses now is measured with a fixed precision and these measured values in a binary form are the final result of the whole pulse code modulation process. We get a series of numbers which still represents the full speech waveform -with some minor errors- and which again can be used to reproduce the analog waveform for presentation of the speech signal through loudspeaker.

The process of pulse code modulation has not wasted anything, but we now can process speech samples by number crunching techniques in fast digital signal processing systems. There are many well known operations to get more intimate knowledge about the information within the speech signal. Some of the more important basic operations are transforms, correlation, and prediction. As an example of a very special speech-oriented operation pitch analysis will be described.

## 2.1 Transform Operations

Since almost 200 years the fundamental principle of signal transform is well known by the Fourier transform. This transform represents the signal by description through harmonic waves, the Sine and Cosine waves. We call the result a Fourier spectrum. Fig. 2 gives an example: The time domain signal in Fig. 2a can by Fourier transform be represented by its power spectrum, where the phase information is lost (for speech intelligibility phase information is not very important).

Speech signals normally are no stationary signals but they change their waveforms within very short sections, lasting in the mean about 20 to 30 ms. If we choose a sample frequency of 8 kHz such a segment or block of 20 ms contains 160 samples of the original speech waveform. This array of 160 samples will be called a vector and all discrete transform operations can be interpreted as operations within an n-dimensional vector space in this case e.g. a 160-dimensional vector space. We call such operations which concentrate only on a well defined short segment of a signal waveform short-time operations. This means, we suppose the speech signal would not change its parameters within this short segment (which is not really true, but the error is small enough).

The principle of a signal transformation can be easily understood by the vector operation shown in Fig. 5. Here only a two-dimensional signal space is shown. The signal vector $\underline{X}^T$ is described by its two components $(x_1, x_2)$, but the mathematical principles are always valid for higher dimensional signal spaces too. The task of the transformation is to transform the basic set of values into a new transformed space which would be better adapted to the characteristics of the signal.
The original n-dimensional signal vector

$$\underline{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad (1)$$

should be transformed into a new vector $\underline{Y}$ by a linear operation

$$\underline{Y} = \underline{A}^T(\underline{X} - \underline{\mu}) \qquad (2)$$

Here $\underline{A}^T$ is a transformation matrix whose column vectors are the basis functions of the new coordinate system. The vector $\mu$ adds a shifting operation by which the centering of the new coordinate system could be further enhanced in respect to the signal vectors $\underline{X}$. Now the new coordinates $(y_1, \tilde{y}_2)$ are much better suited to describe the original vectors with smaller numbers. The value range of a quantizer for such a transformed signal can therefore be much smaller than that of the original quantizer.

One of the most important transforms is the Fourier transform. Here the new basic vectors are the Sine and Cosine functions. The original speech vector after Fourier transformation is expressed in terms of Sine and Cosine waves. The result is normally called a spectrum or a frequency domain representation of the speech signal. This sort of representation is very advantageous because every linear system like the vocal tract produces harmonic waves, and there is a clear evidence that the human ear makes a frequency analysis.

Fig. 6 shows such a digitally computed speech spectrum of the German word "sieben". The frequency axis ranges to about 4 kHz and the duration of the digit was about 800 ms. A new short time spectrum is computed every 10 ms. Here we can see that the following spectrum differs only slightly from the preceding one. Only when the explosion of the sound /b/ happens we notice a very fast onset of this sound after a pause in which the explosion has been prepared. The spectral energy is marked by the darkness of the discrete points and we can see that e.g. in the case of the sound /I/ there are about three frequency areas with high energy, at about 500 Hz, 2600 Hz and 3100 Hz. The pattern of these formants is relatively constant during the sound. On the other side the formant change from the sound /ə/ to /n/ is quite well marked. Every short-time spectrum has about 100 points and so the frequency distance between neighbouring points is about 40 Hz. This is a frequency distance which normally is comparable to the human ear's frequency selectivity.

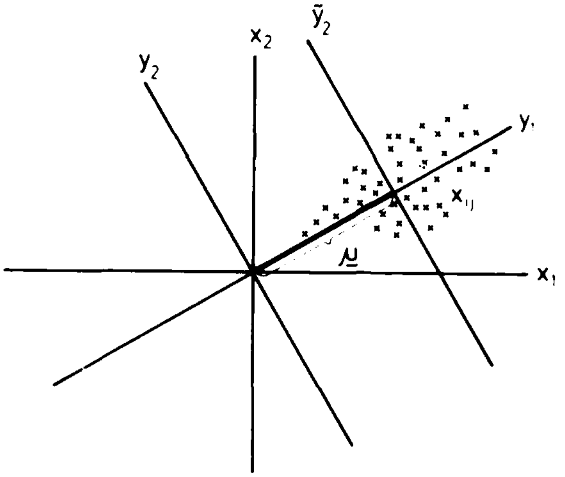


Fig. 5: Principle of transformation in vector domain

In practice the matrix operation from Eq. 2 is done via a very efficient procedure called the Fast Fourier Transform FFT. The number of multiplications necessary is about n•ldn, where n is the number of points. Here we need about 660 multiplications every 10 ms, or one multiplication might last maximally 15 µs. This is quite a long time for modern signal processors which can do this transform in real time. For many applications in speech processing 100 points are too much and so groups of points are joined to make a more rough spectral analysis, a digital variant of the long known bandfilter analysis.

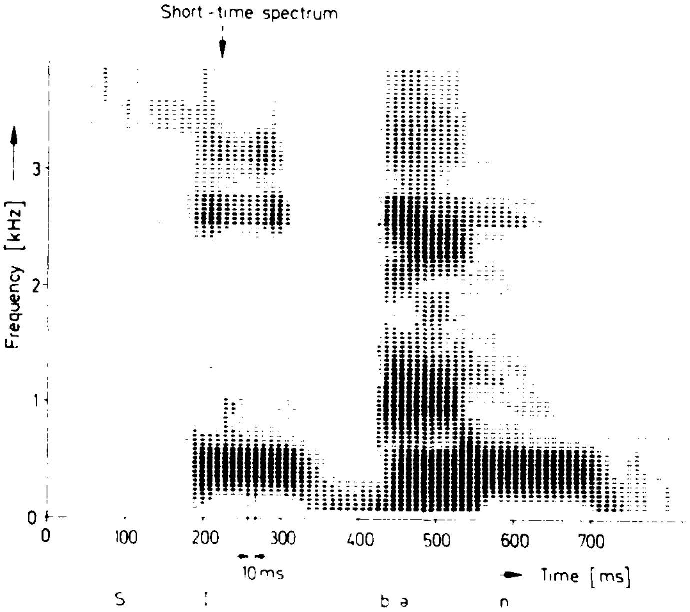


Fig. 6: Spectral pattern of the German word "Sieben" (engl. "seven").

## 2.2 Autocorrelation

We have seen that speech signals are linear superpositions of harmonic waves. This means at the same time that consecutive samples are highly correlated, a speech curve is not stochastically jumping. But far beyond this fact there are still periodicities in the signal which result from the periodic excitation of voiced sounds. Such periodicities can be easily detected by autocorrelation. Fig. 7 shows the principle. The speech samples x(m) -here we prefer not to use the vector writing- are delayed for a varying number of samples k and the delayed and non-delayed signal are multiplied to form the autocorrelation function

$$\phi(K) = \sum_{m} x(m) \cdot x(m+K) \qquad (3)$$

where n is the number of samples which the speech segment contains.

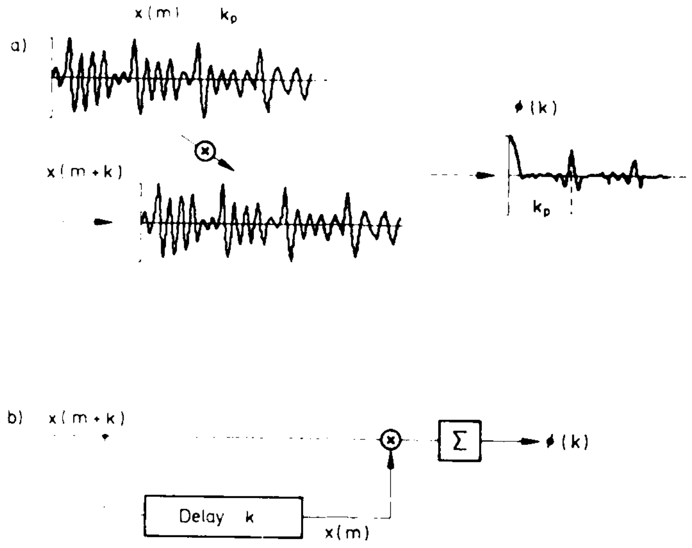


Fig. 7: Principle of autocorrelation analysis.

Of course we can do this operation only within a short-time segment because the speech characteristics change. The correlation function $\phi(k)$ shown in Fig. 7a is near 1 at very small values of k. This means that neighbouring samples are quite similar. The next peak marks the periodicity of this voiced speech signal. The value $k_p$ of the pitch period can be easily found by peak picking.

The autocorrelation function is narrowly related to the power spectrum of a signal. The Fourier transform of the power spectrum is the autocorrelation function. The autocorrelation function is still a time-domain function and therefore it gives information concerning the time domain characteristics of the speech signal.

## 2.3 Linear Prediction /3/

Linear prediction is based on the autocorrelation characteristics of a signal. High correlation values $\phi(k)$ mean that on the average a sample $x_i$ is very similar in its value to a sample $x_j$ where the number $k = (j-i)$. So in the mean it is possible to estimate the value $x_j$ from the preceding value $x_i$.

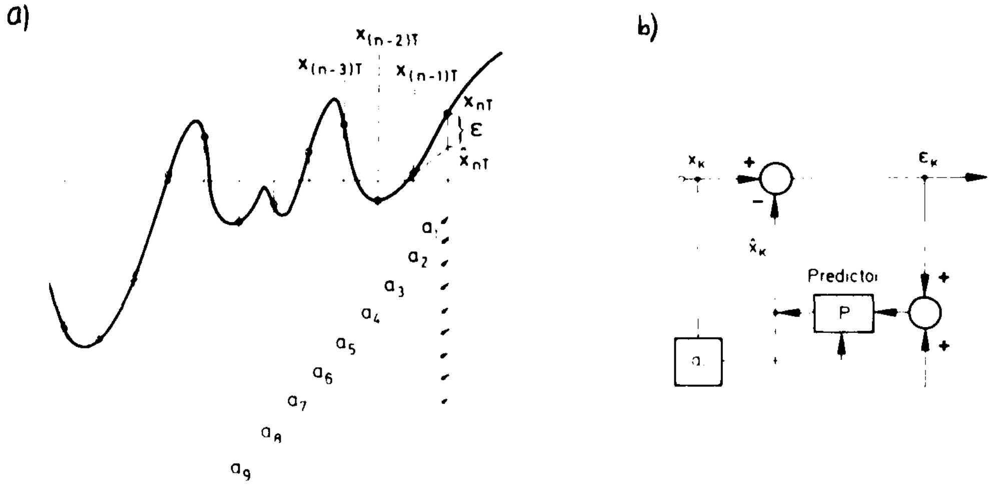


Fig. 8: Linear prediction of signals
a) Principle
b) Recursive prediction scheme

The estimated value of the sample x(nT) in Fig. 8, where T is the sampling period, is

$$\hat{x}(nT) = \sum_{i} a_i \cdot x[(n-i)T] \qquad (4)$$

The $a_i$ are called the predictor coefficients and they are computed from the autocorrelation function by minimization of the predictor error between the estimated and the real value of the signal:

$$\{\varepsilon^2\} = \{(x(nT) - \sum_{i} a_i \cdot x[(n-i)T])^2\} \overset{!}{=} Min \qquad (5)$$

Eq. 5 leads to an algorithm for calculation of the predictor coefficients by a set of linear equations. We can write this again in vector form

$$\underline{M} \quad \underline{a} = \underline{s} \qquad (6)$$

where $\underline{M}$ is a matrix consisting of all the averaged products x(n-i)•x(m-i), $\underline{a}$ is the vector of the predictor coefficients $a_i$ and $\underline{s}$ is the vector of the cornelation coefficients x(n)•x(n-i). The scheme of such a prediction system in Fig. 8b shows that the estimated signal $\hat{x}$ has to be subtracted from the original signal. The prediction error $\varepsilon$ then is minimal if the predictor coefficients are well adapted to the original signal.

In Fig. 9 the original speech signal and the resulting error signal $\varepsilon$ are shown. It can be seen that the error is maximal when the excitation pulse starts a new pitch period. In this moment the free oszillation of air in the vocal tract is interrupted and the prediction fails.

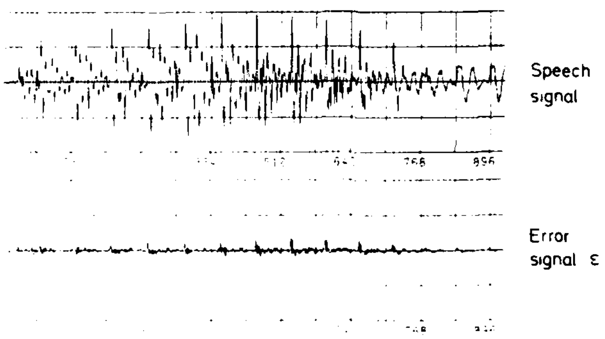


Fig. 9: Speech signal and error in linear prediction.

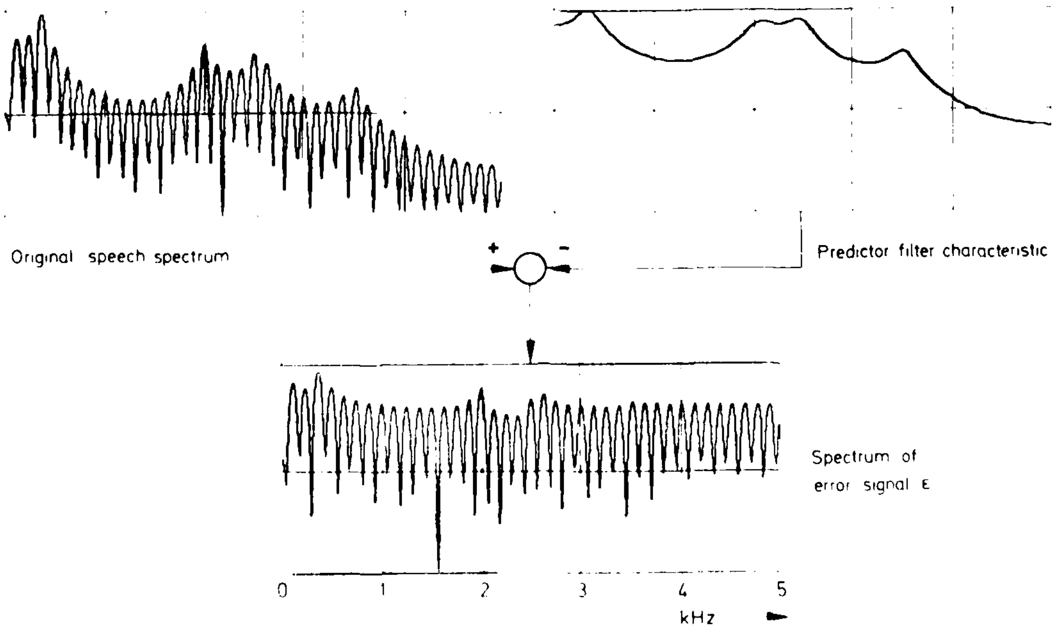


Fig. 10: Inverse filtering of a speech spectrum.

In the spectral domain Fig. 10 interprets linear prediction as a process of inverse filtering. The transfer characteristic of the predictor filter is in a least square sense adapted to the envelope of the speech spectrum. The line structure of such a voiced speech signal can only be reconstructed by predictors with very many coefficients, at least 100 coefficients. For the reconstruction of the spectral envelope like that in Fig. 10 we need about 10 to 14 coefficients. A different sort of predictor for such a periodic structure would be a comb filter. That is a predictor with only few coefficients but the delay between the used speech sample is equivalent to the periodicity predicted. Because the periodicity in the human voice changes in a relatively fast manner it will be necessary to control the delay time of such a comb filter predictor adaptively, and therefore it is necessary to know the exact value of the pitch period.

## 2.4 Pitch Analysis /6/

The algorithms for pitch analysis described should only be representatives for the more complex signal processing techniques which are called feature extraction techniques in Fig. 3. Such algorithms are often not only based on strict mathematical operations but also on some empirically defined rules. Fig. 11 shows first two examples of preprocessing the speech signal for pitch analysis. The first one is the Autocorrelation Function (ACF) already treated in chap. 2.1 and the second is the Average Magnitude Difference Function AMDF which is a sort of simplified autocorrelation avoiding the multiplication

$$\mathrm{AMDF}\,(k) = \sum_{m} \left| X(m) - X(m+k) \right| \qquad (7)$$

This equation is quite similar to equation (3). The most important difference lies in the fact that the ACF has a maximum at its best periodicity value k and the AMDF has a minimum at this point (besides the fact that AMDF only has positive values).

Fig. 11 shows different examples of voiced speech signals and their resulting ACF and AMDF. Both functions are only computed for values around the expected pitch period, not for very small values of k and not for very large ones. Small values of k correspond to high pitch frequencies and vice versa.

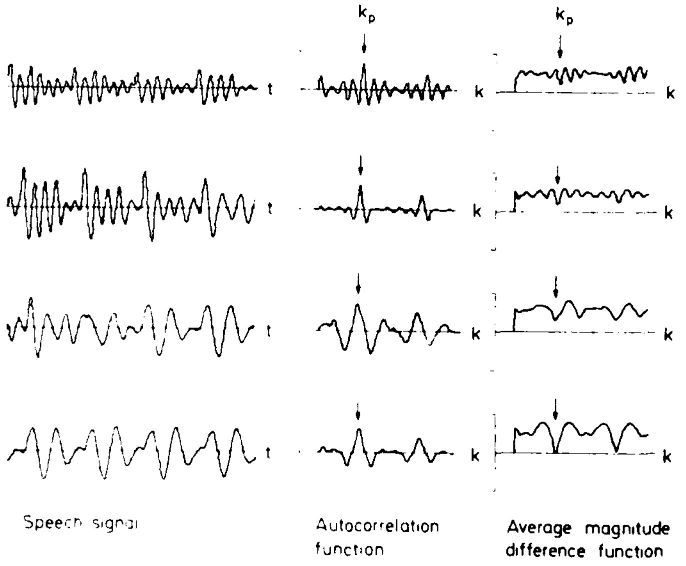


Fig. 11: Periodicity analysis based on Autocorrelation Function ACF or Average Magnitude Difference Function AMDF.

In the case of autocorrelation analysis the first peak is well detectable and enough different from the next peaks which correspond to other frequencies but which are not the real pitch. The AMDF does not show such a clear distinction between the first and the second minimum. Pitch errors could be possible more easily.

To avoid pitch errors which in some speech coders can destroy speech quality a logic postprocessing is necessary. The basic principle is to use a probabili stic model which can learn from the history of pitch contours of the special speakers using the system. So the area for searching maximum or minimum can be restricted and the often possible octave jumps which double or half the original pitch can be avoided.

There are many additional processing stages necessary if e.g. the speech signal is distorted or heavily band-limited. The principal strategy to detect periodicities always uses a sort of autocorrelation or its variants like AMDF.

## 3. SYSTEMS FOR DIGITAL SPEECH TRANSMISSION

Long term research in digital speech has led to a multiplicity of different techniques for speech coding which all are based on the principal algorithms described in chapter 2 but which of course possess many specialities. Most of these systems only have scientific value. Therefore we will describe only those systems which have a real practical significance.

### 3.1 Pulse Code Modulation /2/

This is the most important and the oldest method to code and transmit speech signals. The basic scheme in Fig. 12 is quite simple. The sampled signal is quantized as it has been described already in Fig. 4 Usual data for sampling rate is 8 kHz corresponding to a voice frequency bandwidth of 4 kHz and the quantization is done with 8 bit/sample. So the resulting data rate is 64 kb/s. That is quite a high rate which can not be transmitted over normal telephone channels or over HF charnels. The quantization in PCM is done in a logarithmic manner, small signal values are quantized more precisely than larger values. In this way the signal-to-noise ratio SNR remains constant at a level of about 38 dB for a large dynamic range. This value is better than some degraded analog telephone lines.

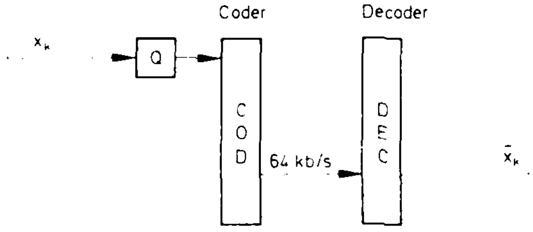


Fig. 12: Pulscode modulation.

### 3.2 Differential Pulscodemodulation DPCM, Deltamodulation /7/

The principal scheme of DPCM is shown in Fig. 13. It is quite similar to Fig. 8 because DPCM needs a predictor which in the most simple version can be a delay for one sample. Then the quantizer Q has to quantize only the difference between consecutive samples. With only slight degradation it is then possible to code speech signals with about 40 kb/s. Every difference sample then is quantized with 5 bits, again in a logarithmic manner.

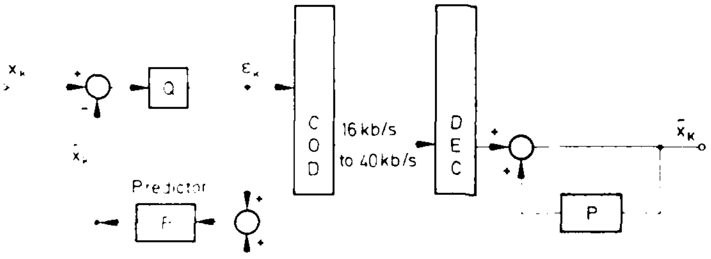


Fig. 13: Differential Pulscodemodulation DPCM or Deltamodulation.

If the data rate of 40 kHz is too high there are further possibilities to reduce the amplitude of the error signal by using a better predictor. This error signal can be quantized with 3 or 4 bits/sample. By further reduction of the speech quality which could only be done in commercial or military applications some 2 bits/sample are still a possible quantizer dimension.

The quality of such a DPCM system can be enhanced by adaptively controlling the predictor coefficients as has been shown in Fig. 8. Such a system is then called Adaptive Differential Pulse Code Modulation ADPCM. The adaptive control can help to make a 16 kb/s system sounding like a 24 kb/s-system but it cannot help to make a well sounding 8 kb/s system.

A slightly different variation of these principles is deltamodulation. The principal scheme of this technique is identical to Fig. 13, but we now use only a 1-Bit-quantizer, which makes hardware very simple. Because a coding with 1 bit/sample is not possible with normal DPCM, we must use a much higher sampling rate. Through this method the differences between consecutive samples will become much smaller, and can so be quantized with 1 bit. There is still another problem: A 1-bit quantizer can only quantize two values, normally 0 and 1, but for speech we need the values -1 and +1 as speech slopes go up and down. Therefore we must leave out the value 0 and the quantizer jumps between -1 and +1. The waveforms of such a delta modulator look like that in Fig. 14a. For very fast signal slopes the delta modulator cannot follow with its fixed step size. The now used Continuously Variable Slope Delta modulator CVSD avoids this drawback by changing the step size of the quantizer. This is in effect similar to changing the predictor parameters.

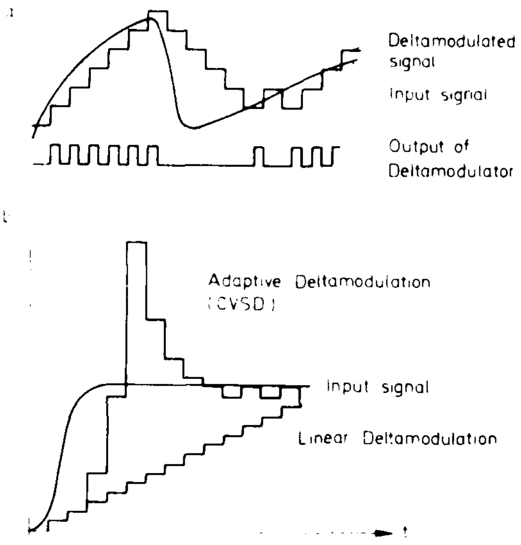


Fig. 14: Deltamodulation signals
a) linear deltamodulation of analog signals
b) linear and adaptive deltamodulation

Fig. 14b shows that such an adaptation can have a faster impulse response than the normal linear deltamodulation. During the last years much more sophisticated methods have been developed to code the error singal in an adaptive way. This means to code and recreate a differential signal like that in Fig. 9 but to transmit only very few parameters. All the methods used are in principle similar: The error signal consists of periodic peaks and in between there is some signal which looks like noise, but is not only noise. Therefore a spectral analysis of this error or residual signal is done, the most important spectral components are coded and transmitted and at the receiver the residuum might be reconstructed. The most important task is to keep the periodicity structure as in Fig. 10 undestroyed. The basic scheme of such a system is shown in Fig. 15. Because now a synthetic error signal is constructed, this can be done by using some information from the original signal too. Therefore the analysis can be done with information from the original and the error signal /8, 11/.

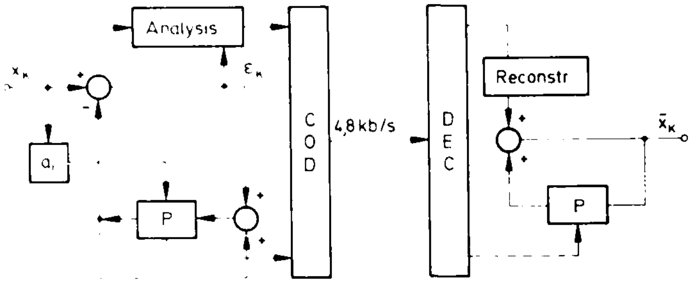


Fig. 15: *Baseband or residual coder*

## 3.3 Transform Coding /9, 10/

Contrary to predictive coding which is operating in the time domain, transform coding does the important operations in the frequency domain. Fig. 16 shows the basic scheme. A set of samples x is transformed into a spectral domain. Besides the well known Fourier transform a much simpler but equally efficient transform has been introduced, the discrete Cosine Transform DCT. The basis vectors of this Transform have some similarity with the Cosine functions from the Fourier transform, but are in their exact shape quite different. These "cosine" functions have much similarity with speech signals and so a representation of speech with these functions is very efficient.

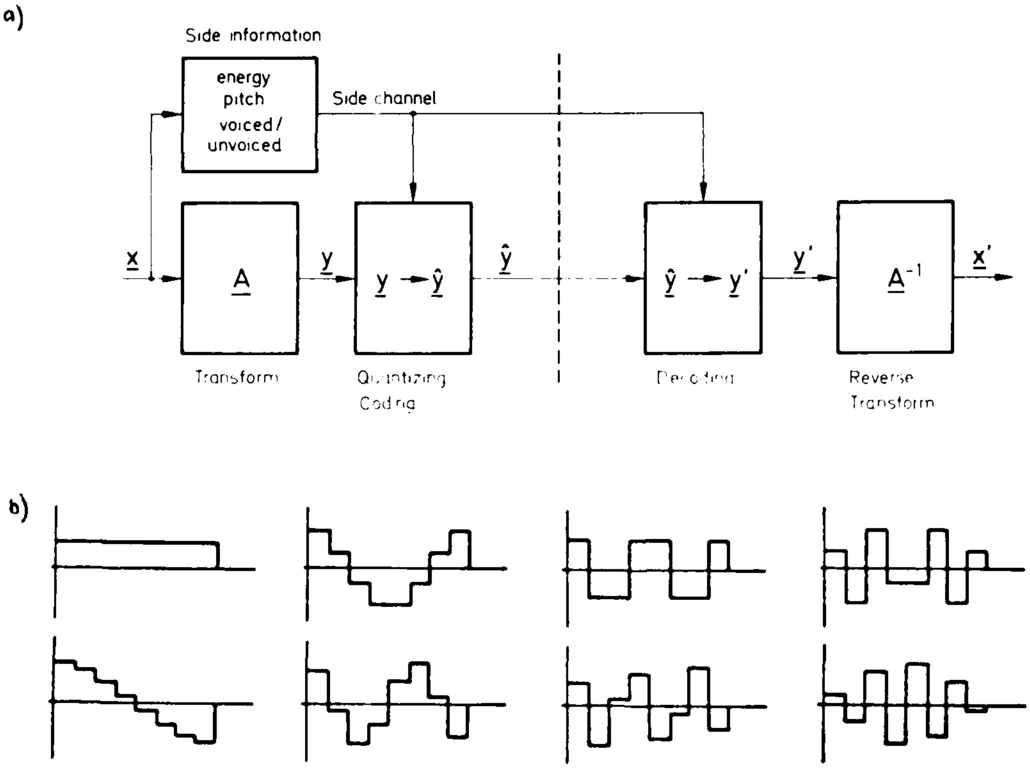


Fig. 16: Transform coding
a) Basic scheme of a transform coder
b) Basic functions for an 8-function Cosine transform.

The sampled speech signal vector $\underline{x}$ has to be transformed blockwise into the vector $\underline{y}$ which in a very sophisticated manner now must be coded and transmitted. Within the receiver both operations are done in a reverse way to reproduce a signal as natural as possible. To take a good block length we use again the same aspects which have already been important at predictive coding. A block should not be much longer than the stationary phase of a speech sound. For a normal articulation rate this would be about 30 ms.

The most important gain in transform coding results from the fact that it is possible to quantize the spectral lines with quite different numbers of bits according to the variance of these lines. To estimate the variances of the different spectral lines it is necessary to get an averaged spectrum of the block to be coded. This can be done like the scheme in Fig. 17. Fig. 17a shows an average long term spectrum of speech. A block spectrum normally is quite different as Fig. 17b shows. For this short time spectrum now an averaged spectrum is constructed whose lines can be interpolated to get a realistic estimated spectrum as a basis for bit assignment within this block (Fig. 17c). This operation has to be repeated for every block. For example in blocks with unvoiced speech signals the normal averaged spectrum will rise to higher frequencies and so look quite contrary to the spectrum in Fig. 17.

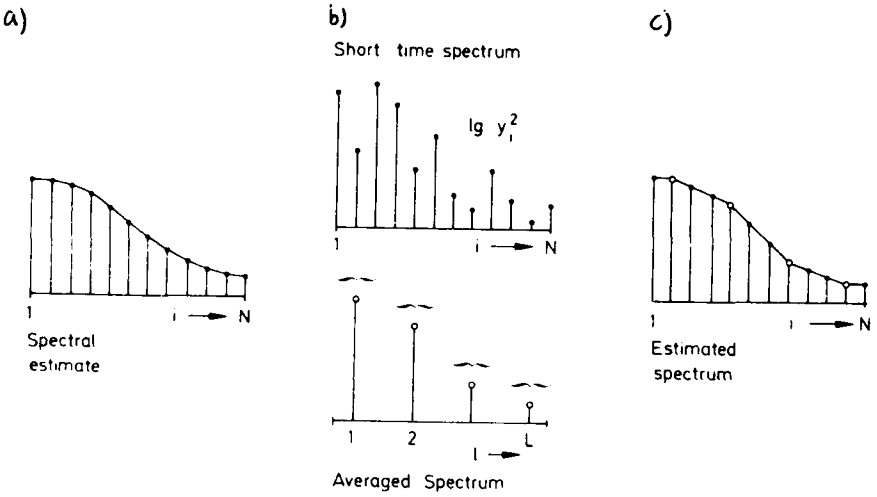


Fig. 17: Estimation of the basis spectrum of a speech block.
a) Long-term averaged speech spectrum
b) Actual block spectrum and averaged spectrum
c) Estimated basis spectrum of the block

The side information additionally necessary has to be transmitted over a special channel. Of course it is also possible to make a much more sophisticated preanalysis of the spectrum to minimize the number of spectral lines which really must be coded. Here e.g. information about the periodicity can be included again.

## 4. ANALYSIS SYNTHESIS TELEPHONY

Fig. 1 has given a principal scheme how the speech signal is produced by the human vocal apparatus. All the operations necessary can be done with digital signal processing too. The exitation function is separated into an impulse and noise function. These produce voiced or voiceless sounds. The three main resonance systems throat cavity, mouth cavity and nose cavity are rather complex mechanical filter systems. There is no principal problem to realize such filters with electronic means. Thus we can build an electronic speech synthesizer but we need to compute the signals for controlling all the parameters which are necessary to produce a naturally sounding and highly intelligible speech signal. These are the pitch frequency to control the pulse frequency of the impulse generator and information about the position of the voiced/unvoiced switch. The control parameters for the articulation cavities can be taken together into a unified filter whose transfer characteristic can be handled in a very flexible manner. The difficulty therefore is not to realize the synthesizer but to get good control parameters and to compute them in real time. Then we can construct an analysis-synthesis system for speech transmission, a vocoder.

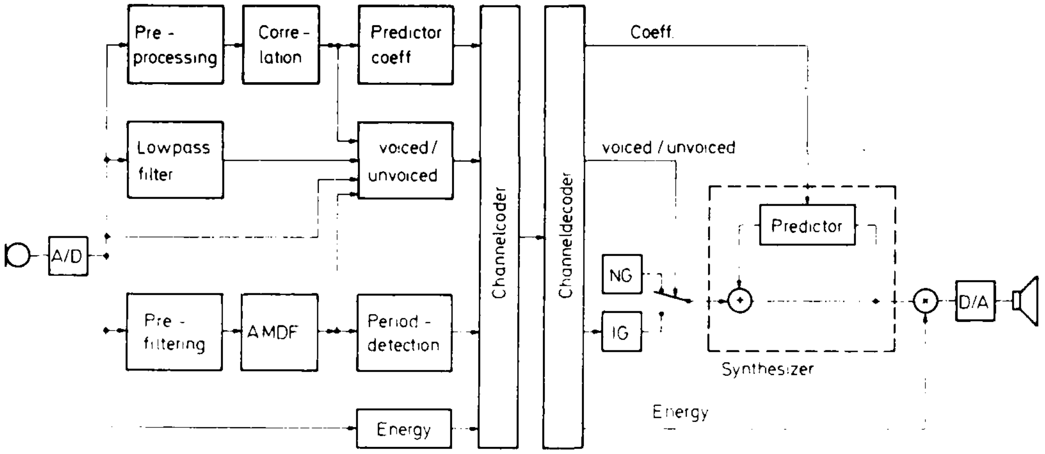


Fig. 18: Linear predictive coder LPC.

These both generators are alternativley switched to the synthesizer filter by a voiced/unvoiced control signal. This excitation signal works like the error signal in an ADPCM coder, but is a quite synthetic signal. The synthesizer filter is a recursive predictor filter whose transfer characteristic can be controlled like that shcwn in Fig. 10. Some resonances can be produced which modulate the flat spectral envelope of the excitation signal's spectrum. So the resulting speech spectrum has the usual formants. The last stage in synthesis makes an adaptive control of speech energy before the digital-to-analog converter remakes the analog signal.

The analyzer needs much more operations. The calculation of the predictor coefficients is done as already described in chapter 2. The same is with pitch detection. The definition of voiced/unvoiced signal must include very different parameters. So the first correla- tion coefficients are as well included as the low pass filtered original signal and the zero crossings of the original signal. All these parameters say something about the spectral content. High low-pass energy means that more low frequencies are within the signal and the probability is high that there is a voiced sound. Otherwise a high zero crossing rate can mean that the signal is of high-frequency content and could be an unvoiced signal. At last the AMDF function is still used whose maximum to minimum ratio gives some hint or there is a voiced or unvoiced sound.

To realize such an LPC vocoder with a universal signal processor makes very fast digital technology necessary because every second some hundred thousand multiplications and adds are necessary for the analysis part and the synthesis part. All these calculations have to be done with at least 16 bit accuracy. A first model of such a vocoder is shwon in Fig. 19. It can transmit speech with a bit rat of 2400 b/s, a data rate that can be transmitted over practically all today existing communication channels. Vocoders are in military use for encrypting the digital bit stream. Analog speech signals cannot be encrypted, they only can be scrambled, a technique by which secure voice transmission is not possible. Because the speech quality of such LPC vocoders is quite good, they will receive wide acceptance in the next years for commercial and military use.

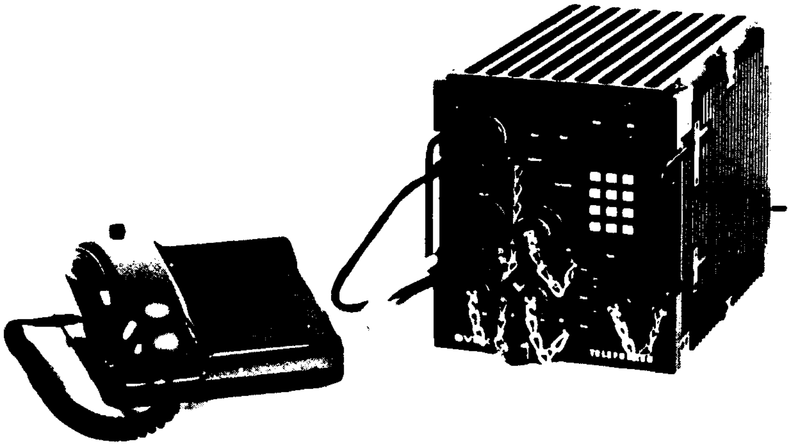


Fig. 19: Hardware realization of an LPC vocoder terminal.

## 5. SPEECH OUTPUT SYSTEMS /13/

The LPC vocoder has shown that it is possible to produce high quality speech signals with electronic means. Therefore vocoders became important not only for speech transmission but for speech output from computers and for voice messaging where speech signals have to be stored and reproduced on demand. The most simple systems for speech output are announcement systems. In the last years very flexible inquiry systems have been introduced where people can get information via telephone, e.g. about railway or airline departures. The speech signals which have to be produced in such a system can be based on prestored words or sentences or the system can create quite new speech signals from basic knowledge about speech production.
The first sort of speech output systems are half-synthetic /14/. Such systems consist of the the blocks shown in Fig. 20.

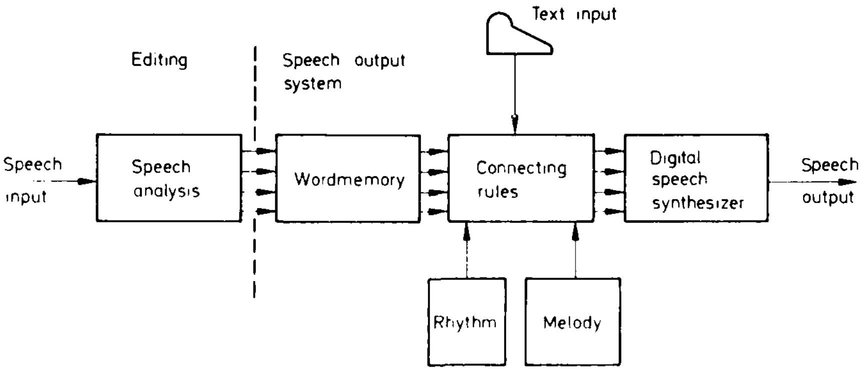


Fig. 20: Segment based half-synthetic speech output system.

The text which should be spoken has first to be analyzed. This is done not directly with the speech analysis system but by a text analyzer. Words and phrases which should be combined from the stored segments have to be identified and the combinatorial rules which define the later necessary control of prosodic parameters must be fixed. Then the vocabulary has to be spoken and analyzed on its LPC parameters. These parameters can be stored and the quality of the speech might be directly tested by synthesizing the intended speech signals. Especially the combinations have to be tested to verify the naturalness of rhythm and melody in the final system. The speech synthesizer today is always an LPC synthesizer. Former used channel vocoders or analog speech concatenators cannot produce speech with a high quality.

The editing stage must not be connected directly to the system as in Fig. 20, but sometimes it is practical if there is a possibility to change the vocabulary through the user and so it would be necessary to integrate some new words into the system.

Such half-synthetic speech output systems are often used as public announcement systems or for special commercial announcement like airport information systems for airline pilots. The information can be transmitted via telephone or radio channels. Sometimes it is necessary to have simultaneously many output channels for many customers with quite different announcements. Then a system like that in Fig. 21 can be helpful. The speech segments are stored in a large data base and a multiplexer and control unit with short-time memory composes the intended announcements for the special customer. The customer itself can control the information he wishes through the telephone dial or use speech input where possible. Half-synthetic systems of course have a limited vocabulary which can only be changed with some effort. Therefore the ultimate speech output systems will be a total synthetic text-to-speech system /15/.

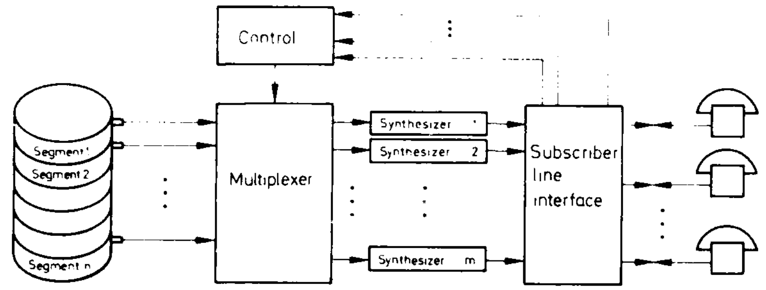


Fig. 21: Multiplex speech output.

The basic structure of a text-to-speech system is shown in Fig. 22. The text input first has to be segmented into its basic elements. This of course is very language dependent. In German e.g. there are a large number of compound words. In English words are only concatenated to build a composite unit. For English it could be satisfying to use a large vocabulary with all the phonetic transcriptions of every word including information about the prosodic parameters like stress or melody, in German this is not possible because stress changes dependent from the word combinations. From this linguistic-phonetic processor we get out a precise description of the articulatory parameters. A human speaker knowing all these agreements should be able to speak the text perfectly even if he would not know the language.

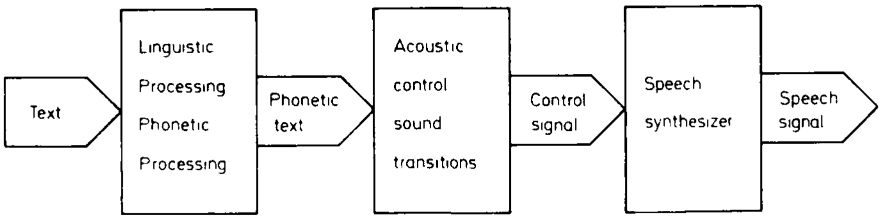

Text
Linguistic Processing Phonetic Processing
Phonetic text
Acoustic control sound transitions
Control signal
Speech synthesizer
Speech signal

Fig. 22: Principle of text-to-speech synthesis

The next stage in the processing knows all the rules for articulation which are implicitly known to the human speaker by a long term use of his articulatory apparatus. This stage knowshow a sound changes if a transition from this sound to a next one has to be made. With all this knowledge this stage calculates the parameters to control the final speech synthesizer which is again an LPC synthesizer, controlled by pitch, voiced/unvoiced and LPC coefficients. A text-to-speech system gives total freedom in vocabulary. The most serious drawback is that it can only be used for one language. But this is a common problem in speech output. In half-synthetic output systems it is possible to concatenate or store very flexibly different languages but it is not quite easily possible to change this vocabulary. In the text-to-speech synthesis systems vocabulary changes are easy but language changes are not possible if the system is not multilingual by its construction /17/. Here much research work has still to be done to develop a really well sounding mutlilingual system.

## 6. FUTURE SYSTEM ASPECTS

Speech coding and transmission will in commercial telecommunication systems more and more be integrated into speech recognition and synthesis systems. This enables not only the normal man-to-man communication but also a flexible integration of data input and output from EDP systems. A simplified version of Fig. 3, with emphasis on the telecommunication and data processing aspect can make this more clear in Fig. 23. One very important aspect of ideas for total voice systems must integrate transmission techniques.

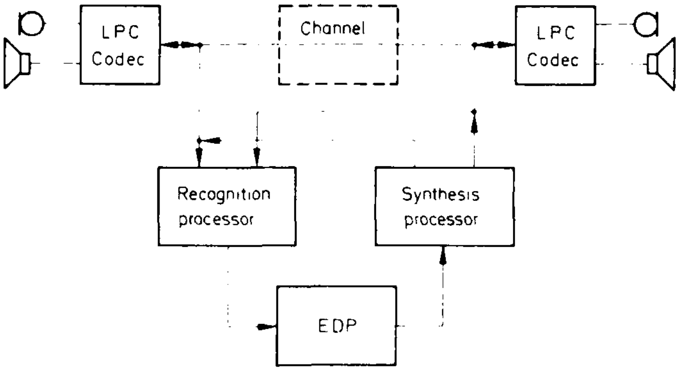


Fig. 23: Speech processing and telecommunication.

Speech coding for transmission has prepared many of the important parameter and feature processing techniques necessary to recognize and synthesize speech signals. Speech coding will in the future too bring deeper knowledge about the important characteristics of speech signals because human judgement about speech quality always is very critical. Speech analysis and synthesis can learn from that.

Another important aspect is that speech coding techniques have prepared efficient digital processing systems working in real time. The same or slightly modified processors can be used in recognition and synthesis of speech. So both techniques can learn and profit from each other to promote the total voice system.

## LITERATURE

/1/ Flanagan, J.L.
Speech Analysis, Synthesis and Perception
Springer Berlin, New York 1972.

/2/ Cattermole, K.W.
Principles of Pulse Code Modulation
Iliffe Books Ltd., London 1969.

/3/ Markel, J.D., Gray, Jr., A.H.
Linear Prediction of Speech
Springer Berlin, New York 1976.

/4/ Rabiner. L.R., Gold, B.
Theory and Application of Digital Signal Processing
Prentice Hall, Inc., Englewood Cliffs, New Jersey 1975.

/5/ Gold, B., Rader, C.M.
Digital Processing of Signals
Mc Graw-Hill, New York 1969.

/6/ Dubnowski, J.N., Schafer, R.W., Rabiner, L.R.
Real-Time Digital Hardware Pitch Detector
IEEE Trans. Acoust., Speech & Signal Proc., ASSP-24, No.1, Feb. 1976, pp.2-8.

/7/ R.E. Crochiere, J.M. Tribolet
Frequency Domain Techniques for Speech Coding
J. Acoust. Soc. Am. 66(6), Dec. 1979, 512-530.

/8/ H. Katterfeldt
A DFT-Based Residual Excited Linear Predictive Coder (RELP) for 4,8 and 9,6 kb/s
ICASSP '81, Atlanta, USA, 827-829.

/9/ R. Zelinski, P. Noll
Adaptive Transform Coding of Speech Signals
IEEE Trans. Acoust., Speech & Signal Proc., Vol. ASSP-25, No.4, Aug. 1977, 299-309

/10/ R. Zelinski
Ein System zur adaptiven Transformationscodierung mit cepstraler Steuerung und Entropiecodierung
FREQUENZ 36 (1982) 7/8, S.193-198.

/11/ Viswanathan, R., Russel, W., Makhoul, J.
Voice-Excited LPC for 9.6 kBPS Speech Transmission
Int. Conf. Acoust., Speech & Signal Proc. 1979, pp. 558-561.

/12/ Tremain, T.E.
The Government Standard Linear Predictive Coding Algorithm: LPC 10
Speech Technology, Vol.1, No.2, April 1982, 40-49.

/13/ Flanagan, J.L.
Computers that Talk and Listen: Man-Machine Communication by Voice
Proc. IEEE, Vol.64, No.4, April 1976, 405-415.

/14/ Mangold, H.
SPRAUS gibt jedem Computer Stimme
Funkschau 4/1981, S.66-70.

/15/ Mangold, H., Stall, D.S.
Principles of Text-Controlled Speech Synthesis with Special Application to German
in L.Bolc, Ed., Speech Communication with Computers
C. Hanser Verlag, München und Wien, MacMillan London 1978, S. 139-181.

/16/ Mangold, H., Schenkel, K.D.
Mensch-Maschine-Kommunikation mit Sprachsignalen
Techn. Mitteilungen PTT 1/82, S. 40-45.

/17/ Carlson, R., Granstrom, B., Hunnicutt, S.
A Multi-Language Text-to-Speech Module
IEEE Int. Conf. on Acoust., Speech & Signal Proc. 1982, pp. 1604-1607.

# TECHNIQUES FOR AUTOMATIC SPEECH RECOGNITION

Roger K. Moore

Royal Signals and Radar Establishment
St. Andrews Rd. Malvern
Worcestershire
U.K.

## SUMMARY

This lecture is intended to provide a brief insight into some of the algorithms that lie behind current automatic speech recognition systems. It is noted that early phonetically based approaches were not particularly successful, due mainly to a lack of appreciation of the problems involved. These problems are summarised, and various recognition techniques are reviewed in the context of the solutions that they provide. It is pointed out that the majority of currently available speech recognition equipments employ a 'whole-word' pattern matching approach which, although relatively simple, has proved to be particularly successful in its ability to recognise speech. It is shown how the concept of 'time-normalisation' plays a central role in this type of recognition process and a family of such algorithms is described in detail. In particular, it is shown how the technique of 'dynamic time warping' is not only capable of providing good performance for isolated word recognition, but how it may also be extended to the recognition of connected speech (thereby removing one of the most severe limitations of early speech recognition equipment). It is also demonstrated how word sequence information can be used to increase the performance of both isolated and connected word recognisers. Finally, a pair of techniques are presented which address the specific problems faced by systems which are to be used by more than one speaker, or in noisey environments. It is concluded that, although current speech recognition algorithms are still relatively unsophisticated, they nevertheless exhibit a level of performance which can be useful in a wide range of well constrained task environments.

## INTRODUCTION

It is now thirty-one years since the first paper to describe a technique for recognising spoken words was published [1]. Since that time, many different techniques have been proposed, ranging from the ridiculously simple, to the dreadfully complicated. In the early years, researchers typically followed a traditional pattern recognition approach, believing that speech was a highly redundant signal containing a sequence of invariant information bearing elements called phonemes. The classical early speech recogniser thus took the form of a pre-processor, to selectively reduce the amount of data present, a feature extractor, typically to identify formant peaks, a segmentor, to divide the signal into phonemic segments, and a classifier, to recognise the individual phonemes from their features (see figure 1). Discovering which word was spoken was then simply a matter of looking up the sequence of recognised phonemes in a kind of dictionary.

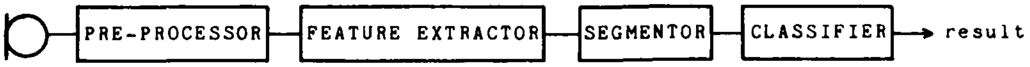


Figure 1: Typical structure of an early automatic speech recogniser.

Schemes of this type abounded in the fifties and sixties but, for reasons which should be apparent from Dr. Hunt's lecture on 'The Speech Signal', they were all doomed to failure.

The reasons why automatic speech recognition is not such a straightforward endeavour as one might imagine may be summarised under four main problem areas:

First, the speech signal is normally continuous, that is to say, there are no pauses between the words in a spoken sentence, nor are there any other acoustic markers which identify where the word boundaries might be. For example, figure 2 shows a speech spectrogram of the phrase "we were away a year ago"; the only pause in this sentence is the middle of the "g" in "ago"! Consequently, techniques for recognising speech automatically must be somehow able to spot words embedded within a surrounding sentence.

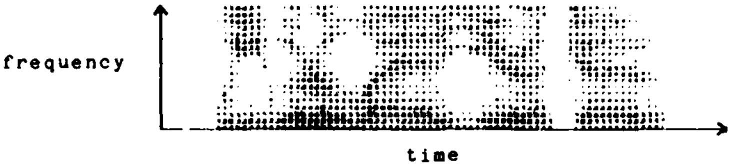


Figure 2:  Speech spectrogram of the phrase "we were away a year ago".

Second, speech signals are highly variable. One person's voice is quite different to another's due to differences in age, sex or accent. Even for a given speaker, his voice will be different on different occasions; sometimes he will speak loudly, sometimes softly, sometimes a whisper, or he might speak fast or slow, or he might even have a cold or be tense. All these factors, and more, may affect a person's voice. In fact, even if a person tries very hard, it is virtually impossible for him to say the same word in exactly the same way on two different occasions. For example, figure 3 shows the word "helicopter" spoken three times by the same speaker; note how the patterns are similar, but not identical. Also, since speech is continuous, adjacent words affect each other to the extent that their beginnings and ends can change quite significantly. For example, the phrase "bread and butter", if spoken quickly, may become "bread'n butter", or "breb'm butter" or even "bre'm butter"! The problems of variablity can therefore be characterised as those conditions which cause speech patterns which one would like to be the same to in fact be quite different. Consequently, one requires techniques which are capable of dealing with patterns which are similar, but not identical.

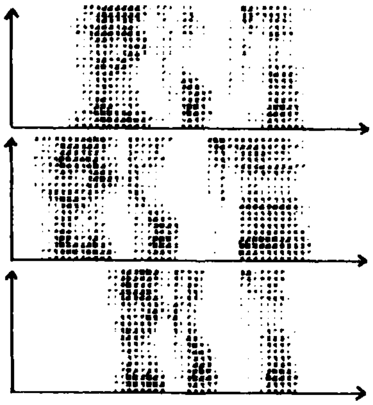


Figure 3:  Three spectrograms of the word "helicopter".

The third problem area is ambiguity. This is characterised by those conditions whereby patterns which one would like to be different, end up looking the same. For example, there is no acoustic difference between "to", "two" and "too". Similarly, "grey tape" sounds exactly the same as "great ape"! The implication here is that one needs techniques which are able to decide on the identity of a particular word after first taking into account the identities of the surrounding words.

The fourth problem area results from the fact that the speech signal is, of course, a part of the complex system of human language. Consequently, it is often the intention behind a message that is more important than the message itself. That is, one might want a system to correctly understand a message, rather than recognise each individual word accurately. For example, the most useful answer to the question "Can you tell me the time?" is "10.15" not "Yes, I can". Therefore, an advanced speech recogniser would be expected to incorporate techniques which would enable it to use the meanings of words in order to interpret what has been said.

This lecture is going to concentrate on techniques which have been found to be particularly successful for tackling the first two problem areas, namely continuity and variability. Techniques in the other areas are the subject of current research, and have not yet found their way into commercial products. The techniques which will be presented here have found practical use, but they too are the subject of continuing research. We are still a long way from having all the answers to any of the problems described above.

## ISOLATED WORD RECOGNITION

In order to make automatic speech recognition a practical reality, it is first necessary to overcome the continuity problem. The technique for solving this is very simple; tell the speaker that he must put artificial pauses between his words, thereby sacrificing naturalness in favour of greatly simplifying the recognition process. Since the positions of the words in such a sentence can now be determined fairly easily, it is then just a question of recognising each word individually. This technique became known as 'isolated word recognition', and machines that use the technique are called 'isolated word recognisers'.

It has already been pointed out that the phonetic approach to speech recognition is too difficult at present, hence most successful techniques for isolated word recognition use the following principle to recognise the individual words:- A word to be recognised is compared with a set of pre-stored reference words (often called 'templates'), and whichever stored word is found to be most similar to the unknown word determines the recognition result. The scheme is referred to as 'whole word pattern matching'. Figure 4 illustrates the idea; a pre-processor turns the speech waveform into some other useful representation (such as a sequence of spectra, or LPC coefficients), a segmentor isolates each word by using the silences between them (a technique known as 'endpoint detection' [2]), and then a comparison module compares the unknown words with each of the templates, and outputs the results. Before, anyone can use such a recogniser, it first has to be given the reference templates, and this process is known as 'training the machine'. Each word is spoken in turn, passing through the pre-processor and the segmentor in the same way as for recognition, and then the individual reference word patterns are stored away inside the machine.

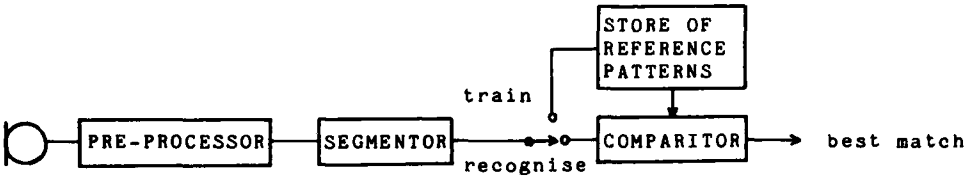


Figure 4: Structure of a typical isolated word recogniser.

Such a machine will only work if the pattern for a word to be recognised is sufficiently similar to the reference pattern for the same word inside the machine. However, it has already been pointed out that the variability in speech is such that this might not be the case. Hence to overcome a major variability problem, the differences between speakers, it is usual for such recognisers to be trained on a single speaker: the person who intends to use the machine. For the same reasons, performance is best if the user trains the machine immediately before he intends to use it. Such systems are referred to as being 'speaker-dependent'.

The key to the success of this simple approach to speech recognition lies in the comparison process. It should already be obvious that an absolute comparison cannot be used, but that some sort of correlation process is required. However, even this is not sufficient since, having eliminated speaker variability by using only one speaker, the major outstanding source of variability is that the same word is very rarely the same length on different occasions. For example, in figure 3 it can be seen that the three versions of the word "helicopter" all have different lengths. Consequently, the patterns which need to be compared may be different sizes, and this is a problem for a simple correlation technique.

The solution, therefore, is to 'time-normalise' each word such that all words have the same length. In practice, the timescale of a particular word is treated as if it were made of rubber, and the pattern is stretched or compressed to the standard length. In the simplest schemes this is done uniformly along the length of the pattern such that if a word has to be doubled in length, each part of the word is doubled. Hence this technique is known as 'linear time-normalisation'.

Figure 5 illustrates the process on a pair of utterances of the word "helicopter". The two original patterns are shown at right angles to each other so that the two timescales can be compared. It is clear that the vertical utterance is much longer than the horizontal one. The rectangle on the right is prescribed by the lengths of the two words, and the diagonal line is the linear time normalisation relationship between the two. The third pattern is the result of stretching the horizontal one to the same length as the vertical one. It can be seen that the two vertical patterns are more similar than the two original patterns, hence the usefulness of the technique.

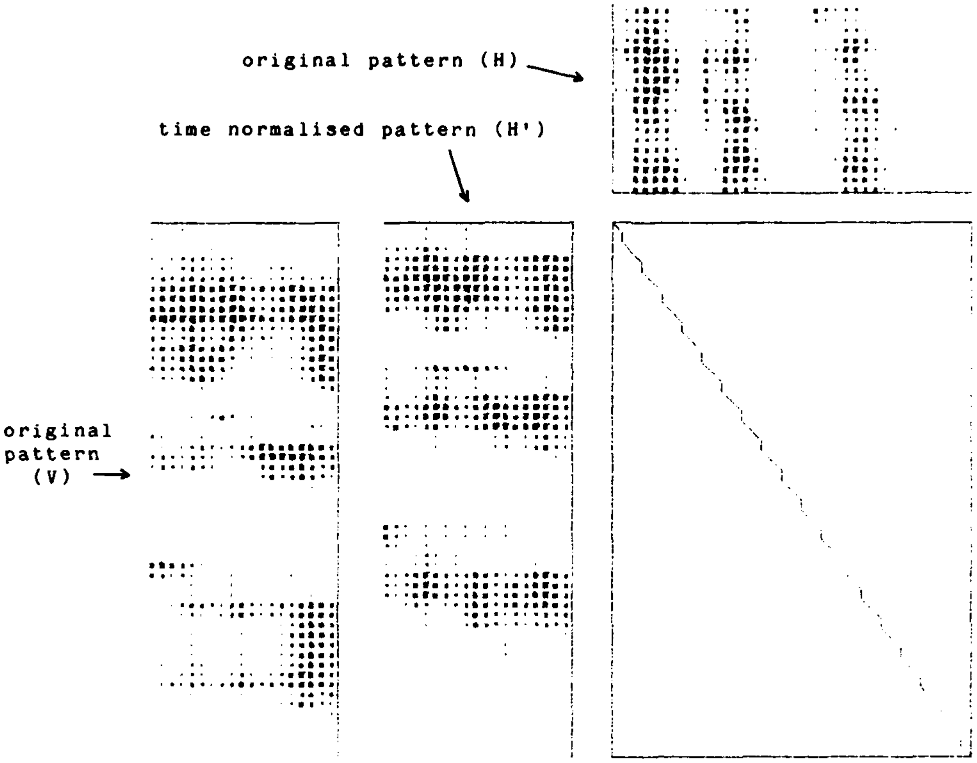


Figure 5: Demonstration of linear time-normalisation.

To calculate the actual similarity, it is common to compare each of the individual frames, or in this case spectra, using some kind of distance calculation, and then to sum these over the entire pattern. So for two speech patterns V (vertical) and H (horizontal) a similarity measure might take the form:

$$D = \sum_{i=1}^{I} [ \sum_{j=1}^{J} ( V(i,j) - H'(i,j) )^2 ]$$

where $H'$ is the time-normalised version of H, I is the length of V and H' in frames, J is the number of parameters in each frame, and D is the distance between the two patterns. If D is zero, then the time-normalised patterns are identical. Typically, I might be chosen such that the normalised patterns are 1/2 second long, and J, for a filter bank pre-processor, might be somewhere between 8 and 20 channels. For LPC parameters it would be usual to employ the Itakura metric in place of the sums of squares to calculate the distance between frames [3].

A number of commercial isolated-word recognisers have been produced which incorporate these techniques. However, individual machines will not be reviewed here, the intention is merely to give an overview of the principles involved.

The performance of the algorithm can be quite useful if the number of words which the machine has to distinguish between (the 'vocabulary') is kept small, 10 to 30 words for example. For the ten digits "zero" to "nine", one could expect recognition accuracies up to about 97% under ideal conditions. The actual performance obtained will depend, amongst other things, on the consistency of the speakers, the exact nature of the pre-processing, and the number of training examples allowed per word. This level of performance, whilst not perfect, has proved suffiently good to allow machines of this type to be used in fairly simple applications, examples of which will be described in later lectures.

For larger vocabularies, the recognition accuracy obtained using linear time-normalisation can drop significantly, so low in fact that practical use is out of the question. The reason for this is that linear normalisation is not a very gooJ rodel of what happens when people make words longer or shorter. In practice, what actually happens is that some sounds are changed more than others. For example, if you listen to yourself say the word "three", first fast, and then slow, you can hear that the "-ee" changes length more than the other sounds. This effect is apparent in figure 5. Although linear time-normalisation has made the patterns the same length, it has still not made them particularly similar to each other. By eye one can see that the patterns have similar structures, and one can imagine that by distorting the timescale of the horizontal utterance non-linearly, it could be made much more like the vertical utterance. Figure 6 shows exactly this, the line in the rectangle is no longer linear and the horizontal pattern is distorted accordingly. This result was achieved by a person deciding which parts of the word required lengthening and which parts needed shortening.

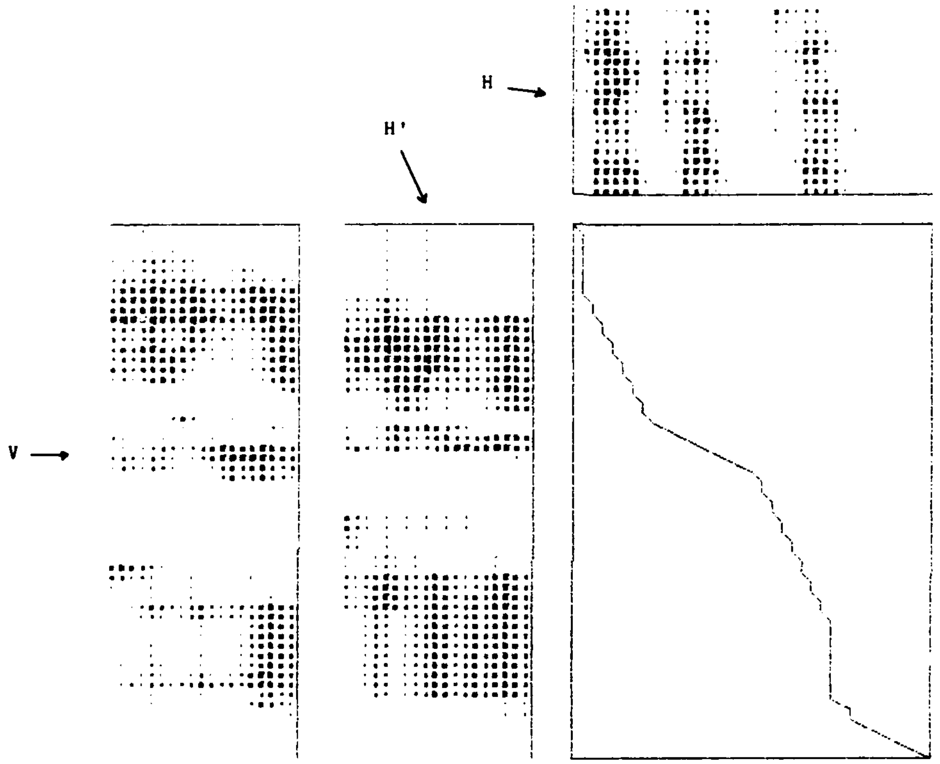


Figure 6: Demonstration of non-linear time-normalisation.

This technique is known as 'non-linear time-normalisation', and it can be seen that, by improving the comparison process, the performance of an isolated-word recogniser may be raised.

Of course it is necessary to find the non-linear distortion automatically, rather than by hand (as in figure 6) and this presents a rather difficult computational problem. Obviously, there are many millions of possible distortions, that is, there are many possible lines across the rectangle between the two timescales. However, rather than search all the possible distortions in turn (potentially a very time consuming process) it is possible to apply the mathematic technique of 'dynamic programming'. Figure 7 shows the result of using dynamic programming on this particular pair of words. Note how similar the original vertical utterance is to the non-linearly distorted version of the horizontal distance. Since dynamic programming is guaranteed to find the best possible distortion, this result is 'optimal non-linear time normalisation'.

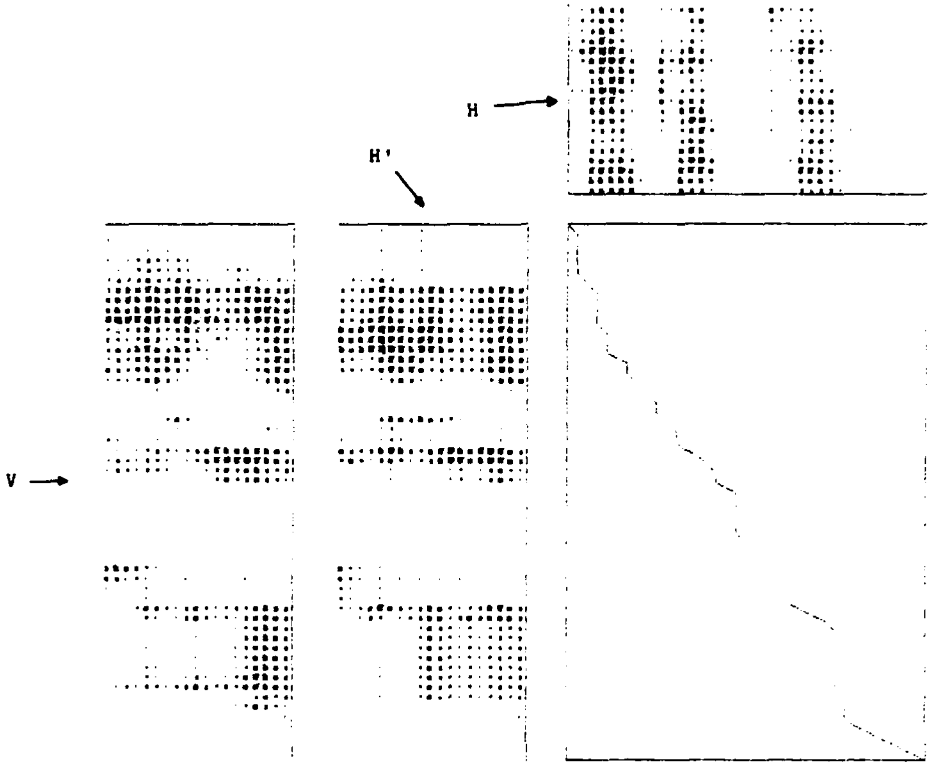


Figure 7: Optimal non-linear time-normalisation using dynamic programming

Optimal non-linear time-normalisation, or 'dynamic time warping' (DTW) as it has become known, is still computationally quite expensive in comparison with linear time-normalisation, but it is still the most efficient way of getting the required answer, and the result is guaranteed to be the best. Once the distortion has been made, then a distance between the two time-normalised patterns may be calculated as described earlier. The actual technique of DTW will be described later in the lecture.

To illustrate how the technique is used in practice, figure 8 shows an example of isolated word recognition using dynamic time warping. In the example there are three reference patterns, the digits "one", "two" and "three", shown vertically. The horizontal utterance is the word to be recognised, actually a "two". The unknown word is compared with the three reference patterns using the DTW technique, and the resulting three non-linear time distortions are shown. Also shown are the numbers which are the distances between the unknown and each of the reference patterns. The best match is determined by the smallest distance (the highest similarity). Hence the unknown word is recognised correctly as "two".

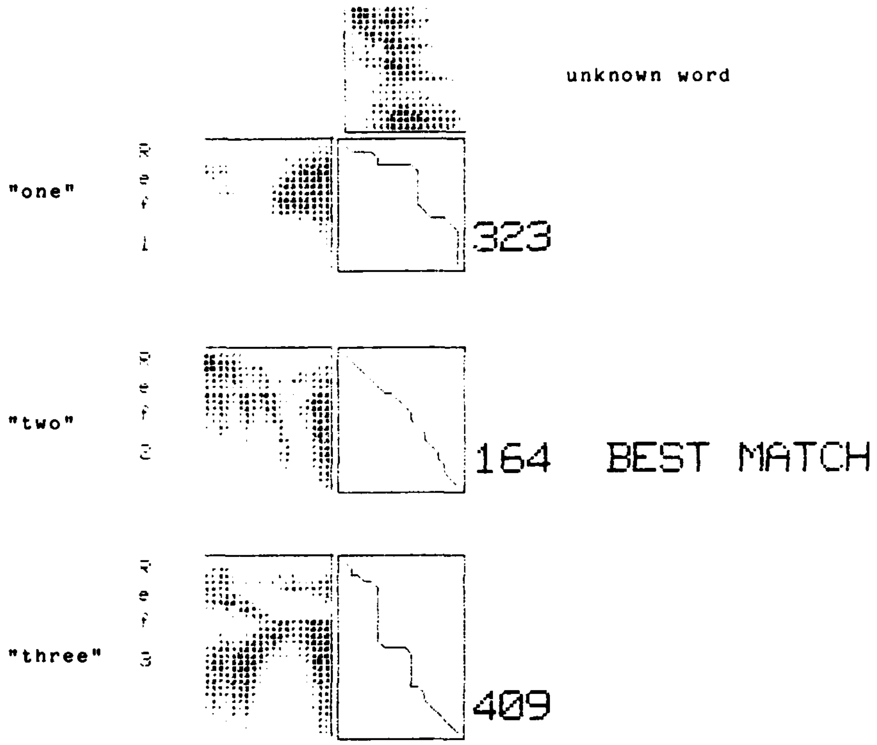


Figure 8: *Isolated word recognition using dynamic time warping.*

To interpret the non-linear distortions, it should be noted that when matching two words which are the same, such as in figure 7 and the correct match in figure 8, the distortions tend to be subtle non-linear variations on a linear theme. On the other hand, when two words are different, such as the two incorrect matches in figure 8, the distortions tend to be grossly non-linear. This is because it takes a very severe distortion of the timescales of two different words to make them even remotely similar.

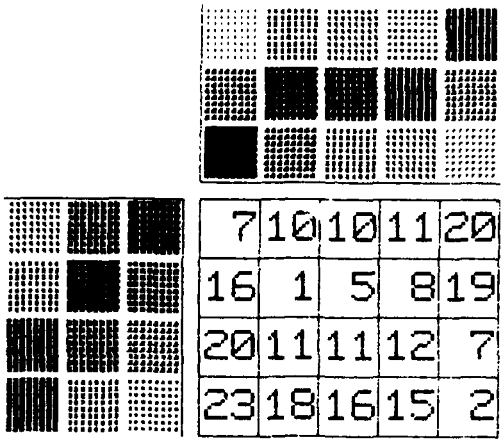
In practice it is possible to have more than one reference pattern per word. This enables more variability in pronunciation to be captured and the performance will be improved. Similarly, some training procedures involve averaging different examples to obtain a suitable reference pattern. The Bell laboratories 'robust' training procedure is a hybrid of the two, combining averaging with a statistical clustering procedure [4].

Since the dynamic time warping technique is able to provide a far more realistic compensation process than linear time-normalisation, the performance of isolated word recognisers based on DTW is significantly better. Greater variability (in length) can be accomodated, hence larger vocabularies are possible. Typically, for the ten digits, one could expect recognition accuracies greater than 99% (remembering that we are still talking about speaker dependent isolated-word machines).

## DYNAMIC TIME WARPING

As has already been stated, dynamic time warping is based around the mathematical technique of dynamic programming (DP). This technique is a sequential optimisation process whereby many local optimisation decisions are combined in order to find a globally optimal solution to a problem. The process, as it applies to dynamic time warping, can be readily understood with reference to a few diagrams.

Dynamic time warping is essentially a two-stage process. Figure 9 illustrates the first stage. Two abstract speech patterns are shown, one vertically and one horizontally. Each pattern has time frames consisting of three parameter channels, the vertical pattern has four frames, the horizontal has five. The matrix in the centre is known as the 'distance matrix' and it contains numbers which correspond to the distances between each frame in one pattern and each frame in the other pattern. For example, the number "20" in the top right hand corner indicates that the first frame of the vertical pattern is quite different to the last frame of the horizontal pattern. Similarly, the "1" in row-2 column-2 indicates that the second frames of each pattern are very similar. The distances are actually calculated by taking the sum of the squares of the differences in each parameter channel for each pair of frames.

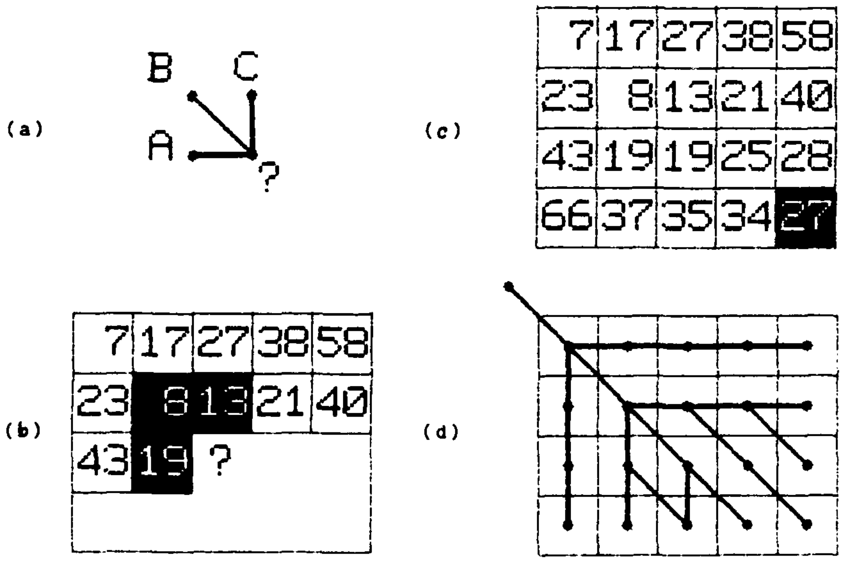


| | | | | |
|---|---|---|---|---|
| 7 | 10 | 10 | 11 | 20 |
| 16 | 1 | 5 | 8 | 19 |
| 20 | 11 | 11 | 12 | 7 |
| 23 | 18 | 16 | 15 | 2 |

Figure 9: Dynamic time warping: distance matrix.

The creation of the distance matrix is thus the first stage. The second stage is to find a 'path' through the distance matrix from the top left hand corner to the bottom right hand corner which has, along its length, the minimum sum of distances. This path is the required non-linear relationship between the two timescales for these patterns. In other words, the basic function of dynamic time warping is to find the least-cost distortion of two patterns in order to make them look like each other.

The procedure for finding the best path out of all the possible paths is where the dynamic programming comes in, and it involves the successive application of a 'local decision function' to the distance matrix in order to construct a 'cumulative distance matrix'. Figure 10 illustrates the process.



| | | | | |
|---|---|---|---|---|
| 7 | 17 | 27 | 38 | 58 |
| 23 | 8 | 13 | 21 | 40 |
| 43 | 19 | 19 | 25 | 28 |
| 66 | 37 | 35 | 34 | 27 |

| | | | | |
|---|---|---|---|---|
| 7 | 17 | 27 | 38 | 58 |
| 23 | 8 | 13 | 21 | 40 |
| 43 | 19 | ? | | |
| | | | | |

Figure 10: a) Local decision function, b) partially filled cumulative distance matrix, c) completed cumulative distance matrix, and d) decison matrix.

The local decision function is shown in figure 10(a). This is a three way decision function which says that a path may arrive at any particular point either vertically, diagonally or horizontally. So, for any point in the cumulative distance matrix, the smallest cost of getting to that point is the minimum of the costs of getting to the three previous points. However, it is also necessary to take into account the cost of being at a particular point in the first place, and that is the number in the corresponding place in the distance matrix (figure 9).

Figure 10(b) shows the cumulative distance matrix in the process of being filled in. The "?" indicates the point being considered, and the three previous points are highlighted. The cost of getting to the point is the minimum of 19, 8 or 13, and the cost of being at that point is 11 (from the distance matrix). Hence the number entered into the cumulative distance matrix is 19 (8+11).

Figure 10(c) shows the cumulative distance matrix completely filled in. The number in the bottom right hand corner is highlighted because this is the overall distance between the two patterns. This is the number which is shown in figure 8; it is the sum of distances along the least-cost path through the distance matrix. To find the path it is necessary to remember, at each point in the calculation of the cumulative distance matrix, exactly which local decisions were made (horizontal, vertical or diagonal). Figure 10(d) shows all of these decisions, and it can be seen that they form a tree radiating from the top left hand corner (this is where the calculation started). The actual minimum cost path is obtained by tracing back along the local decisions starting at the bottom right hand corner.

Referring back to the distance matrix, figure 9, the calculation shows that the least-cost path takes the route 7+1+5+12+2, and it can be seen that no other path has a sum lower than 27.

To summarise, the formulation for the distance between two speech patterns obtained using dynamic time warping is based upon the following recursive expression:

$$D(i,j) = \min\,[D(i-1,j)\ ,\ D(i,j-1)\ ,\ D(i-1,j-1)] + \sum_{k=1}^{K} [V(i,k) - H(j,k)]^2$$

where $1 \leq i \leq I$, and I is the number of frames in speech pattern V, $1 \leq j \leq J$, and J is the number of frames in H, and K is the number of parameters per frame. The overall distance between the two patterns V and H is D(I,J).

## CONNECTED WORD RECOGNITION

The importance of DTW lies in two areas. First, recognition accuracy is much greater than with linear time-normalisation. Second, it has in fact provided a rather neat solution to the continuity problem. It has turned out to be possible to extend the technique from isolated to connected words using a relatively simple modification to the algorithm. Conceptually, the modification can be understood as follows:- In the isolated word situation, DTW is able to find all the non-linear temporal relationships (paths) between the unknown pattern and the reference patterns. Figure 8 shows three paths: the best path (for the correct match) and two sub-optimal paths. The best path explains the relationship between the unknown word and one of the reference patterns. To recognise an unknown sequence of connected words, therefore, it would be necessary to find a path which explains the relationship between the unknown phrase and a sequence of reference patterns. In practice this proves to be fairly easy, it is merely necessary to allow paths to jump from reference pattern to reference pattern whilst computing the dynamic time warping. The trajectory of the best path then determines the recognition result.

Figure 11 illustrates this technique quite clearly. The reference patterns are the same words as in figure 8, but this time the unknown pattern consists of a sequence of words (actually "11213"). The best path, determined by DTW, is shown, and it can be seen to be jumping around from reference pattern to reference pattern. The trajectory reveals that the phrase is recognised correctly as "11213".

There are a number of variations on this particular technique [5,6], but the very simple, yet very effective, implementation described here is attributed to John Bridle [7].

The technique represents a new and exciting development in automatic speech recognition since the restriction of using isolated words may be removed. As a consequence there are now a number of connected word recognisers available commercially, and as a group they are the most advanced machines around.

The potential for more natural communication with machines is obviously higher with connected word recognisers, but it is worth remembering that they are also speaker dependent, and perhaps more importantly, they do not take into account the variations which may occur at word boundaries, as described earlier. This is because the technique assumes that a connected phrase consists of a sequence of isolated words with little modification, hence such machines will not be able to recognise the "and" in "bre'm butter".

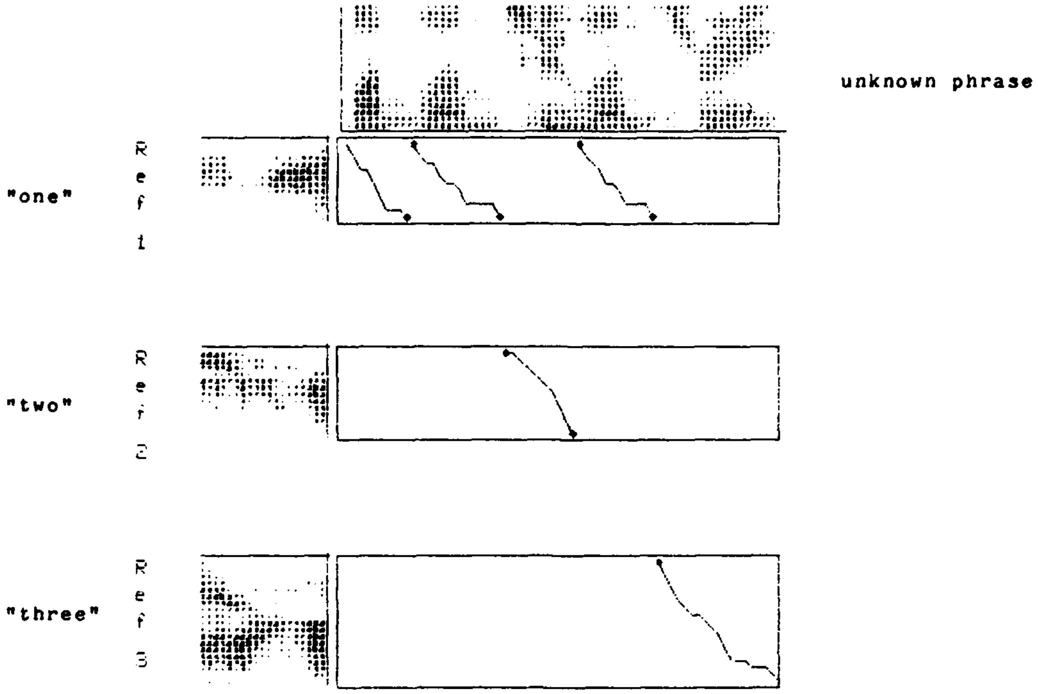


Figure 11: Connected word recognition.

Therefore, in order to achieve good recognition accuracy for connected speech, it is necessary to ask the operator to speak as clearly as possible, and not to run his words together too much. It is also common to train such a machine on reference words which are spoken fairly abruptly (as in figure 11), since otherwise there may be length differences which are too great for the DTW to handle. Another scheme is to train on word sequences in order to include word boundary modifications in the reference patterns [8]. This is usually a bootstrapping procedure whereby the connected word recognition algorithm itself is used to extract reference patterns from a carrier phrase using, in the first instance, normal isolated references. This technique is known as 'embedded training'. Some other schemes use the same extraction principle, but train on each possible word pair sequence.

Of course a connected word recogniser may be used to recognise isolated words, and the performance is just the same as a TW based isolated word recogniser.

## SYNTAX

A limiting factor on the performance of both isolated and connected word recognisers is the size of the vocabulary they use. In general, the more words there are in the vocabulary the worse the performance will be (due to variability). Consequently a popular technique for maintaining high performance with large vocabularies is to exploit the fact that in most tasks not every word can follow every other word. In other words, a syntax (a grammar) may be used to limit the alternative words to be considered by a recogniser at each point in a sentence. For example, in a sentence such as "hello victor tango two this is ...." the active vocabulary in a recogniser may be cut down to just the military alphabet in order to recognise the next word.

There are a number of ways of specifying a syntax, but the most popular is in the form of a state transition diagram. Figure 12 illustrates a syntax for a voice controlled calculator. It can be seen that the diagram describes sentences such as "what is two plus four compute" and "put nine times alpha into beta compute". The overall vocabulary size is 23, but the maximum number of words that need to be considered at any point is 14, and in some places only one word is allowed. The average number of legal words is 8 and this is known as the 'branching factor' of the syntax, the lower the branching factor the higher the performance.

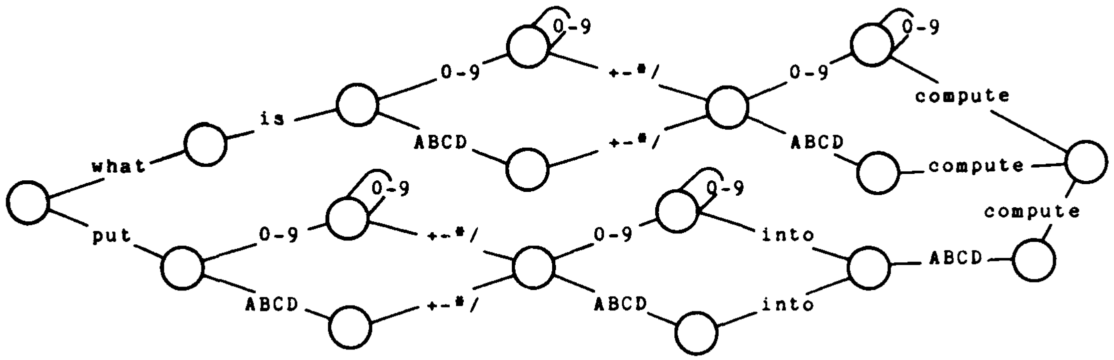


Figure 12: Syntax for a voice controlled calculator.

The implementation of syntax is very easy for isolated word recognisers; most machines have facilities for specifying which reference patterns are to be considered during a recognition match. For connected word recognisers the DTW process may be modified on the basis of a state transition diagram to only allow the path to jump between reference patterns if such a jump is legal in the syntax. Hence, the syntax can be made an integral part of the optimisation process and connected word recognisers with this facility are able to find the best syntactically valid interpretation of a connected utterance.

The only problem using syntax with a speech recogniser is that the user has to remember the allowable sequences of words. If he says a word which is syntactically illegal, then the recogniser may be forced to misrecognise it, even if the word is in the overall vocabulary.

## MULTIPLE SPEAKERS

It has been pointed out several times that all of the techniques presented so far are speaker dependent, due primarily to the use of speaker-specific reference patterns. In general, performance for a person using somebody else's reference patterns is pretty poor, the exact level of performance being dependent on the similarity between the two people's voices. Of course speaker dependent systems may be used by any number of users, as long as they each train the machine first. If they do this each time they use it, then they will get better performance than if they do it once and recall those reference patterns on later occasions. However, if the vocabulary is large, the training may be too tedious to do more than once.

Neither of tnese techniques is suitable for the situation where the users are unknown and thus will not have trained the machine at all (such as a person making an enquiry over a telephone). In this instance the most successful technique has been based on selecting representative reference patterns from a range of speakers sufficiently wide to cover the pronunciation differences of the expected users [9]. It has been found, using the robust training technique on data from fifty male and fifty female speakers, that between 6 and 12 reference patterns per word gives good performance over a wide range of unknown speakers. Of course there is a limit to how far this procedure can be applied. Eventually, with a large cross-section of accents, problems of ambiguity arise because the pronunciations of different words begin to overlap. Nevertheless, if the accent variations in an expected user population are relatively small, then the technique can be quite useful.

## NOISE

In most environments noise is always present, and this is another source of variability when it comes to recognising speech. The effects of noise on a recogniser are threefold, first, the segmentor may make errors in determining when speech is present, second, noisey speech is more likely to be misrecognised, and third, the speaker may change his vocal characteristics because of the noisey environment.

There are a number of techniques which can be used to combat the effects of noise, but first it is worth pointing out that, whatever the situation, higher performance is almost always obtained if training is done in the environment in which a machine is going to be used. This is particulary true in a noisey environment.

Obviously a noise cancelling microphone helps considerably to overcome background noise, since a less noisey speech signal then reaches the recogniser. It is also possible to use a separate piece of noise cancelling equipment between the microphone and the recogniser. Alternatively, noise compensation may be integrated directly into the recognition algorithm itself [10]. In particular, the frame to frame comparison process in the DTW may be modified to take into account an estimate of the effects that the noise has on the individual parameters in the frames, hence recognition proceeds by actively ignoring data which is known to be noisey.

If the noise is impulsive, rather than continuous, then it is sometimes possible to train the recogniser on these sounds, and then allow it to recognise them as they occur. This technique has been found to be particularly successful in coping with breathing noises!

## CONCLUSION

This lecture has provided a brief overview of a number of techniques which are central to the operation and application of practical automatic speech recognition equipment. Many of the algorithms are relatively simple in concept, and very few of the many problems facing automatic speech recognisers have been satisfactorily solved. Nevertheless, the techniques are such that machines are now available which display a level of performance which is suitable for many limited applications [11].

## REFERENCES

[1] K H Davis, R Biddulph and S Balashek. "Automatic recognition of spoken digits". J. Acoust. Soc. Amer., vol.24, 1952, pp 637-642.

[2] L R Rabiner and M R Sambur. "An algorithm for determining the endpoints of isolated utterances". Bell Syst. Tech. J., vol.54, 1975, pp 297-315.

[3] F Itakura. "Minimum prediction residual principle applied to speech recognition". IEEE Trans. Acoust. Speech, Signal Processing, vol.23, 1975, pp 67-72.

[4] L R Rabiner and J G Wilpon. "A simplified, robust training procedure for speaker trained isolated word recognition". J. Acoust. Soc. Amer., vol.68, 1980, pp 1271-1276.

[5] H Sakoe. "Two-level DP matching - a dynamic programming based pattern matching algorithm for connected word recognition". IEEE Trans. Acoust. Speech, Signal Processing, vol.27, 1979, pp 588-595.

[6] C S Myers and L R Rabiner. "Connected digit recognition using a level building DTW algorithm". IEEE Trans. Acoust. Speech, Signal Processing, vol.29, 1981, pp 284-297.

[7] J S Bridle and M D Brown. "Connected word recognition using whole word templates". Proc. Inst. Acoust., Autumn 1979.

[8] L R Rabiner, A Bergh and J G Wilpon. "An improved training procedure for connected-digit recognition". Bell Syst. Tech. J., vol.61, 1982, pp 981-1001.

[9] L R Rabiner, S E Levinson, A E Rosenberg and J G Wilpon. "Speaker-independent recognition of isolated words using clustering techniques". IEEE Trans. Acoust. Speech, Signal Processing, vol.27, 1979, pp 336-349.

[10] D H Klatt. "A digital filter bank for spectral matching". Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, 1976, pp 573-576.

[11] L R Rabiner and S E Levinson. "Isolated and connected word recognition - theory and selected applications". IEEE Trans. Communications, vol.29, 1981, pp 621-659.

# Speaker Differences in Speech and Speaker Recognition

*Melvyn J. Hunt*

National Research Council of Canada
National Aeronautical Establishment
U66, Montreal Road
Ottawa, Ontario
K1A 0R6
Canada

## Summary

This talk is concerned with the differences between speakers. The range of ways in which speakers differ is surveyed, with distinctions being drawn on the one hand between physiological and .usage differences and on the other hand between those differences stemming from the larynx and those stemming from the vocal tract. Methods of dealing with speaker differences in speaker-independent and speaker-adaptive speech recognition systems are discussed. This is followed by a discussion of the exploitation of speaker differences in speaker recognition systems. The latter discussion is divided into a consideration of speaker verification, in which a speaker is trying to prove his identity, and speaker identification, in which the identity of an unknown speaker has to be discovered and the speaker cannot be expected to cooperate in producing a predetermined phrase. The talk is concluded by a summary of the present state of the art in dealing with speaker differences together with some guesses about the prospects for practical systems in the near future.

## Introduction

This session is concerned with the differences between the speech of speakers of the same language. I want to look at what sort of differences there might be, and how those differences might be useful in some tasks where the aim is to determine the identity of the speaker, and a nuisance in others where the aim is automatic recognition of what is being said irrespective of who is saying it.

### What sort of differences are there?

One way in which speakers differ is in the choice of the words and expressions they use. Humans probably make use of such information in recognizing each other, but I see little possibility of any automatic system having that capability in the near future, so I propose to leave this kind of speaker difference out of the discussion.

To consider how differences in voices can arise, it is useful to go back to the *source/filter* model of speech production that I talked about in the earlier session. You may remember that the production of voiced speech can be quite accurately modeled as an acoustic source, the larynx, feeding into a linear filter, the vocal tract. Both the source and filter contain speaker-dependent characteristics, these characteristics depending in both cases partly on physiology and partly on usage.

#### Excitation Differences

The larynx varies in size between individuals; in particular, a man's larynx is much larger than a woman's and consequently average values of adult male fundamental frequency (around 100Hz) are much lower than corresponding values for adult females (around 200Hz).

Fundamental frequency is also to some extent under the control of the speaker. Differences in speakers' use of fundamental frequency seem to be to be describable in two ways: we can describe them in an absolute sense, for example by measuring the variance of fundamental frequency about its mean value, or we can describe them with reference to the associated sentence structure by noting such phenomena as a tendency for fundamental frequency to rise or fall at the occurrence of a particular syntactic feature in a sentence. I suspect that the latter kind of description is much more important in distinguishing speakers, and the former much less important, than we normally imagine. Evidence for this contention comes from some work in language discrimination carried out by Maidment [1]. He had a group of subjects listen to the fundamental frequency patterns of sentences taken from conversations some of which were in English and some in French. The signal was taken electrically directly from the larynx, and the vocal tract had essentially no influence on it. French and English intonation patterns normally sound quite different, yet in trying to identify which language they were hearing from these intonation patterns with their usual accompaniment of vocal-tract information removed most subjects performed rather poorly (average 63% correct, against a chance level of 50%).

The details of the action of the vocal cords within each excitatory cycle also vary between speakers, and are the major factor in what is known as *voice quality* [2]. Among others, the adjectives *breathy, harsh* and *creaky* are used to describe laryngeally determined voice qualities. We know relatively little about what causes different voice qualities, or even how to classify them, but it seems probable that, like fundamental frequency, voice quality is partly physiologically determined and partly under the control of the speaker. It is likely that in good acoustic conditions listeners use voice quality information to identify speakers, but it tends to get lost when the speech is distorted as it is by being transmitted over a telephone link for example.

Vocal Tract Differences

As far as speech production is concerned, the most obvious and probably most important physiological difference between vocal tracts is in their length. In particular, women tend to have vocal tracts about 15% shorter than men, and adult female vocal tract resonances are consequently about 15% higher in frequency than corresponding male resonances. Whether length differences result in a simple linear scaling of frequencies is still questioned, but linear scaling does seem to be a reasonable first approximation.

There are certainly other physiological differences in vocal tracts - differences in absorption in tract walls, for example, which would affect resonance bandwidths - but for the most part their effects on speech have not been extensively studied.

Differences in vocal tract usage are of many different kinds. They range from an idiosyncratic pronunciation of a single word by a single individual to a tendency for a whole dialect group to use a particular vocal tract setting in a number of speech sounds. As an example of the second extreme, many speakers from the Birmingham area of England tend to have the back of the tongue slightly raised throughout much of their speech.

Between these two extremes there can be differences in how a particular phoneme is realised. Sometimes, the differences can affect a whole sequence of phonemes. For example, in the speech of many speakers of "standard" European French there are three nasalized vowel phonemes that occur in the words *bon, banc, bain;* in the speech of many French Canadians *bon* is pronounced like standard French *banc, banc* like *bain*, and *bain* is pronounced with a nasalized diphthong not occurring in standard French. Sometimes, phonemes can show context-dependent differences: to take another French Canadian example, the /i/ phoneme occurring in *mite* is pronounced by many French Canadians rather like the English vowel in *bit*, but only when it is followed by a consonant, otherwise it has the tenser, standard French form.

Finally, a particular speaker or group of speakers can have a different system of phonemes from the other speakers of the language: the northern English dialect that I grew up speaking, for example, has only one phoneme for the vowels that occur in *luck* and *look* or *putt* and *put*, so the pairs of words sound identical, whereas standard British English has two phonemes and the pairs of words are distinguishable by their pronunciations.

At this point, I would like to say that I am now going to look at how these various kinds of differences are systematically handled and exploited in automatic speech and speaker recognition systems; but our understanding in this area is not yet that advanced. Speaker variations are dealt with instead by a set of pragmatic approaches, and the most I can suggest is that in considering these approaches one should keep the various sources of variation in mind.

## Speaker differences in speech recognition

Differences between speakers pose a major problem for developers of speech recognition devices intended for use by more than one person. Contributors at the first NATO Advanced Study Institute on Spoken Language Generation and Understanding held at Bonas, France in 1979 were asked to identify the major outstanding problem in speech recognition. Without exception they named the need for speaker independence. In the same year panel members at the IEEE Workshop on Speech Recognition held in Pittsburgh were asked the same question and gave the same unanimous reply. Today, four years later, there are still very few speech recognition devices available that work reliably for anyone other than the person who provided the training material.

The importance of a capacity to accept input from more than one speaker depends, of course, on the application. In a central system being accessed briefly by a large number of military or civilian users speaker independence is essential. On the other hand, it is not particularly important for a device that is to be used for a long period by one person, particularly if the vocabulary is small so that total retraining is not a lengthy process. Pilots could, for example, carry a cassette recording with them that would allow them to retrain a system to their own voice before a flight without having to speak the training material each time. Even in this case, however, research on speaker independence may well prove relevant, because it is likely that in learning how to make a system tolerant to variations between speakers we will also learn how to make them tolerant to changes occurring in the *same* speaker caused, for example, by high accelerations or by high levels of psychological stress.

Before I go on, I should point out the speciousness of a claim that is sometimes made for speaker-dependent devices, namely that they provide a measure of security against unauthorized use. A speaker-dependent device simply makes more recognition errors when used by a speaker who did not train it. A level of accuracy that was unacceptably low for the legitimate user could be perfectly acceptable for a determined unauthorized user. To claim speaker dependence as an advantage seems rather like saying that small, cramped cars have the advantage of being unlikely to be stolen by tall thieves.

How can speaker independence be achieved?

Humans carry out speaker-independent speech recognition all the time. When a stranger starts to speak we usually understand him immediately. If he speaks with an unfamiliar accent and if what he says is not highly predictable from the situation, our recognition is error-prone, but it still enormously better than most artificial recognizers can manage. It may be helpful, then, to ask ourselves how humans manage the task. There seem to me to be three mechanisms that may be involved: a listener may derive acoustic/phonetic features from the speech that are invariant across speakers; he may accept a set of alternative production forms; or he may deduce some general characteristics of the speech he is hearing and use them to adapt his recognition process.There is some dispute about the relative importance of the first and last possibilities [3,4], but it seems likely that all three mechanisms are involved to some degree. The fact that we do a fairly good job of understanding a new speaker immediately suggests that our analysis of the speech signal is good at extracting speaker-independent cues. There is no doubt, however, that our comprehension of a new speaker, particularly one speaking an unfamiliar dialect, does get better after we have heard a few sentences from him and thus had a chance to adapt to his voice. Finally, we must store alternative forms of at least some words in order to handle a case like the word *either*, which has two distinct pronunciations, and the pronunciation that a particular speaker will choose is not predictable from the rest of his speech. As we shall see, automatic speech recognition systems have incorporated all three of these mechanisms to varying extents.

Speaker-invariant acoustic representations

The search for invariants has been carried out in two ways. The first is to assume that the acoustic representation generated by the human ear is one which minimizes speaker differences, and a representation that faithfully copies the ear should consequently provide a degree of speaker independence. Some workers [5,6] have reported evidence tending to confirm this assumption, and at least one connected speech recognition system has achieved speaker independence by carefully modeling neural behavior in the ear together with details of the phonetic characteristics of the language used (the only published account I can find of this work [7] deals with an earlier, isolated-word system).

Alternatively, features in the speech signal that are invariant across speakers can be sought by statistical methods. The simplest form of this approach is to apply suitably chosen weights to the features used in the comparison of the speech to be recognized with the reference forms. Each weight would depend inversely on the measured inter-speaker variability of the corresponding feature. Thus, if the ends of words were found to vary more across speakers than the beginnings and middles they would be given less weight in the comparison process. The same applies to regions of the spectrum: My experience [8], for instance, has been that when comparing power spectra across speakers for equivalent speech sounds the energy profile in the first 300 Hz is much more variable than the rest of the spectrum (presumably because of differing fundamental frequencies), and it therefore helps considerably to reduce the weight given to this portion of the spectrum in the matching process. I believe that the Verbex (formerly Dialog) speech recognition system [9], whose use by the State of Illinois civil service constituted the first large-scale application of a speaker independent system, used variability weighting as a function both of frequency and position within the word. A more sophisticated approach [10] takes linear combinations of features to derive speaker-independent linear discriminant functions.

## Alternative reference forms

Turning to the use of discrete, alternative forms, we have to draw a distinction between two kinds of speech recognition systems. In the first kind, which I will call *non-segmenting* systems the basic reference forms are whole-word templates represented as a sequence of vectors describing the power spectrum and spaced regularly apart in time, typically every ten milliseconds. In the second kind, words are broken down into a sequence of phonetically classified segments of variable duration. In these *segmenting* systems, it is the segments  corresponding very loosely to phonemes in some cases - rather than words that form the set of basic reference units. Although, for what seem to me to be good reasons, the trend in practical recognizers is strongly towards word-based systems, the segmenting approach is undeniably more efficient at representing alternative forms of words when the difference is localized in one part of the word as it is in the first syllable of my *either* example. This is because it is relatively easy for the system developer to represent variable portions of words by constructing branching networks from the segmental reference units. Such a network representation was used, for example, in the *Harpy* system [11] developed as part of the ARPA Speech Understanding Project. It has been suggested [12] that branching within words could be incorporated into non segmenting systems, and at least one group has described experiments in constructing such templates [13]. The task is much less straightforward than in the segmenting case, however.

In practice, the only way that non-segmenting systems have allowed for alternative forms is by having separate whole-word templates for each variant. The multiple templates for each word are usually created by taking examples of the word from a hundred or more speakers and averaging together groups of examples that are found to be similar. The Verbex system mentioned earlier used, I believe, three such templates per word. Recognition systems have been demonstrated at Bell Labs [14] that rely exclusively on multiple templates to obtain successful speaker-independent performance. Typically, six variants are used for each word when the speaker population forms a homogeneous dialect group.

It has sometimes been claimed that multiple templates represent by themselves a satisfactory solution to the problem of speaker differences in speech recognition. While the success of the approach is very impressive, I do not believe that they constitute *the entire* solution for the following reasons. First, the use of $n$ templates for each word means $n$ times more storage and $n$ times more computation, so practical considerations dictate that $n$ should be as small as possible. Second, the classic work of Peterson and Barney [15] showed that a sound (specified in terms of the first two formant frequencies) that would represent one vowel phoneme when produced by one speaker could represent a different vowel phoneme when produced by another speaker. These repeatedly confirmed results suggest to me that in automatic systems in which speaker differences are handled exclusively by multiple templates, the distributions in acoustic space of templates representing different words will inevitably start to overlap as we move to larger vocabularies. If the system makes no attempt to learn something of the characteristics of the current speaker, it will have no way of making reliable recognition decisions in the overlapping regions. Finally, while some differences in the pronunciation of a particular word are definitely discrete, others such as those resulting from differences in vocal tract length  are continuous in nature. To try to represent a continuous range of variation by a few discrete points seems, to say the least, inelegant.

## Speaker adaptation

The potential effectiveness of the third approach to obtaining speaker independence, namely adaptation to the current speaker, depends very much on the intended application. The process inevitably takes time, and if it is to be worthwhile, each new speaker must continuously use a system for a period several times longer than the time needed for useful adaptation. On the other hand, if the expected period of use is quite long and the vocabulary quite small, the time overhead in having each new user re enter the complete vocabulary may not seem unreasonable, and it will almost certainly lead to better performance than the more sophisticated adaptation schemes.

Adaptation material can be collected by having the new speaker utter a predetermined phrase before he starts to use the system, or it can be gathered *on line* in the initial portion of the speaker's use of the system. On line adaptation is much less obtrusive, but its use depends on the system being able to provide a reasonable level of recognition performance before any adaptation has taken place. Moreover, unless careful verification is carried out to ensure that early inputs have been correctly recognized, there is a danger that the system might try to extract adaptation information from incorrectly recognized material with disastrous results.

Turning to how adaptation can be achieved, it seems to me that segmenting systems are at an advantage once again. Systems like *Harpy* with a hundred or so speech sound templates can efficiently update their complete set of templates by having the user utter a suitably chosen phrase containing all the sounds in the inventory [16]. Furui has shown [17] that by taking account of the correlations that exist in speaker differences in different speech sounds it is possible to begin to update speech sound templates before an exa.nple of that speech sound has been given.

In word-based systems one way of achieving on-line adaptation if the vocabulary is small is to update the template for each word as it occurs in the input. The updating can consist of a direct replacement of the old template by the newly received example or of an averaging together of the old template and the new example. The Verbex system mentioned earlier incorporated this kind of adaptation.

A second way in which word-based systems can adapt is to assume that the differences between corresponding speech sounds for two speakers are substantially the same. That is, the difference between one speaker's /i/ phoneme and another speaker's /i/ phoneme is assumed to be similar to the difference in the productions of their /e/ phonemes, for example. Such an assumption is likely to be valid for excitation differences and for differences resulting from physiological characteristics of their vocal tracts, such as their lengths, but it will not in general be true for differences in usage of their vocal tracts. In so far as the assumption is valid, speaker differences can be determined by time-aligning corresponding words from the two speakers so that the aligned power-spectrum vectors correspond to equivalent speech sounds. Any consistent differences across pairs of aligned vectors can then be used to construct a speaker-adapting spectral transformation. In some digit recognition experiments I carried out [18] the use of transformations derived in this way reduced the average recognition error rate by a factor of two after just three digits had been input.

To conclude this section, it seems likely to me that future successful large-vocabulary speaker-independent systems are likely to include an attempt at a speaker-invariant acoustic representation together with multiple versions of at least some words and a capacity for speaker adaptation. The three approaches are not competitors: on the contrary, they are likely to be more effective when they operate in concert.

## Speaker Recognition

Speaker recognition is the positive side of speaker differences. I mean the heading to cover two quite distinct classes of problem: *speaker verification*, in which characteristics of a speaker's voice are used to verify that he is who he claims to be; and *speaker identification*, in which there is an attempt to determine whether some speech to be identified could have been generated by one of the speakers known to the investigators. I want to confine the discussion here to *automatic* methods, and so leave out of account the use of human listeners, with or without the use of spectrograms, sometimes misleadingly referred to as voiceprints.

Despite a fair amount of experimental effort on the two problems, there have been remarkably few practical implementations of speaker verification, and - as far as I can tell - no practical use of automatic speaker identification up to now.

### Speaker Verification

Speaker verification has potential applications in the control of physical access to secure areas and in the control of remote access to sensitive information, such as an ability to confine access to personal bank account information to the account holder. I think that the remote access applications are more interesting because there are few fully automatic alternatives.

Verification is set apart from speaker identification partly by the fact that the comparison process is essentially one-to-one rather than many-to-one. A much more important difference, though, is that the speaker is *cooperative* and can therefore be induced to utter a particular phrase. This utterance can then be compared with a version of the same phrase known to have been produced by the person that the current speaker is claiming to be. In this way, equivalent speech sounds in identical contexts can be compared.

Speaker verification systems have been tested with some success over-telephone links [19,20,21], generally with some attempt at reducing sensitivity to linear distortions.

Although the ability to use text-dependent methods makes verification generally easier than identification, the reliability demanded of a verification system may be much higher. Identification used in the early stages of a police investigation to help reduce a long list of suspects to a shorter list can be tolerant of occasional errors, whereas in a verification system controlling access to a sensitive site a single error could be very damaging. Moreover, verification systems are more prone to attack by deliberate imposters. If the same phrase is always used, an imposter could potentially become proficient at mimicking another speaker's production of the phrase. Alternatively, if he could procure a recording of the other speaker uttering the test phrase, he could fool the system simply by playing the recording. It was presumably for these reasons that in the speaker verification system controlling access to a Texas Instruments' computer room [22] the system permuted the words of the original reference utterance and demanded an unpredictable phrase each time an individual presented himself.

### Speaker Identification

Interest in automatic speaker identification comes from the police, security and intelligence organizations, and also from accident investigators trying to determine, for example, who said what just before a plane crash.

Automatic speaker identification is faced with two problems that combine to make the task particularly difficult. The first is the lack of control over what the speaker says. The second is that in almost every practical application the signal from which the speaker is to be recognized - telephone call, intercepted military radio transmission, or aircraft cockpit recording - must be expected to have suffered a significant degree of distortion.

Let us look first at the implications of lack of control over the text. As we have seen, the great strength of speaker verification is that by controlling the text it is able to compare equivalent speech sounds in equivalent contexts. Automatic speaker identification clearly cannot do this, and I see little hope in the foreseeable future of an automatic system being able to spot instances of a particular speech sound in a speaker-independent manner and on transmission-degraded speech with a useful degree of reliability. Even if a system could be made to spot target sounds most of the time, the occasional misses and false alarms that would inevitably occur would seriously bias the data being gathered.

Given these limitations, almost all attempts at automatic speaker identification have worked by seeking to produce a statistical description of the speech typically in terms of the properties of power spectra computed every centisecond - without any regard to what is being said. Thus, much of what constitutes the difference between two speakers will be blurred over - if a speaker pronounces *luck* like *look* the system will hardly notice. It seems rather like taking a handwritten text and cutting each line into a set of thin vertical slices, each slice a few times thinner than the average letter, and then trying to determine the identity of the writer from the statistical properties of an assorted heap of the slices. The surprising thing is that on undistorted speech such a method works quite well.

The traditional way of producing the statistical description [23] starts by computing a set of statistical parameters - means, variances, etc. - of short-term signal properties - filter-bank channel energies, linear predictor coefficients, etc. The differences are noted in these parameters between speakers compared with their variation between speech samples taken from the same speaker. This provides a primary measure of the usefulness of a parameter in speaker identification. However, such parameters rarely turn out to be statistically independent of each other, so their correlations also have to be taken into account. Given this information, and making certain assumptions about the statistical distributions of the parameters, a linear transformation of the parameters derived from each sample can be computed. Provided the assumptions hold, the transformation provides optimum discrimination between the sample sets belonging to the different speakers. The distances in the transformed parameter space are known as *Mahalanobis distances*, and the process forms part of what is known as *linear discriminant analysis*.

Recently, a couple of experiments have been described that step outside the traditional framework. In the first [24], the distribution of short term signal features (in this case, linear prediction log area ratios) in the speech to be identified is compared non parametrically with the distributions generated by known speakers. In the second [25], the speech of each of the known speakers is modeled by a Markov chain, and the probability that a known speaker could have generated the unknown speech is estimated from the degree to which the speaker's Markov model fits the speech data. It will be interesting to see if either of these approaches leads to practically useful systems.

I would like to look now at the second serious problem encountered in automatic speaker identification, that of transmission degradations. Researchers have tried to find features in the speech signal that have some resistance to these degradations. Perhaps the most resistant of all such features is fundamental frequency, the tendency of the waveform in voiced speech to repeat itself periodically. The repetition rate is clearly unaffected by linear or non linear distortions, and it should remain observable in the presence of moderate amounts of steady noise. As one would expect, then, appropriately chosen algorithms can derive reliable statistics of fundamental frequency even from heavily distorted speech. Since automatic systems cannot determine the syntactic structure of the speech samples they are given, the statistics they derive are necessarily of the absolute kind as described in the first section I argued there that such statistics are not likely to be rich in speaker characterizing information, and it is indeed found to be the case that they are not very effective in speaker identification, particularly with the short samples of speech that one must expect to work with in practical applications. Moreover, fundamental frequency is probably the most mood sensitive feature of the speech signal, and in many situations where speaker identification could be useful the speaker would be unlikely to be in a normal, calm state.

Measures of the behavior of the power spectrum ought to be less mood-sensitive. The power spectrum is also richer in information than fundamental frequency is. The drawback, however, is that most properties of the power spectrum are very sensitive to transmission distortions: there is no point in trying to characterize a speaker by his long-term spectral average, for example, if that spectrum is going to be drastically and unpredictably modified by the transmission process.

Spectrum-shaping linear distortion manifests itself as a set of frequency-dependent additive constants in the log power spectrum. For this reason, measures of the variability of the short-term log power spectrum about its mean have been proposed as robust speaker-characterizing features in the presence of linear distortion [26]. These measures do not, however, have any special resistance to noise or to non-linear distortions, and I know of no speaker identification experiments in which they have been successfully used on speech obtained from real - as opposed to simulated - telephone links including carbon microphones (which cause non-linear distortions). My own experience using real telephone speech [27] has been that such measures are somewhat better than measures of the long-term spectral average, though somewhat worse than statistics of energy peaks in the spectrum. Even the results with energy peaks, though, did not approach a practically useful level of performance.

It seems as though we can have successful text-independent speaker identification on undistorted material, and speaker verification results show us that we could probably have successful text-dependent speaker identification on transmission-distorted material; but effective text-independent speaker identification on transmission-distorted material seems so far to be beyond our grasp. I note that in West Germany dynamic microphones are now being used in public telephones, and I wonder whether a trend towards the elimination of carbon microphones and the increasing use of digital transmission might not mean that the troublesome distortions will be substantially eliminated before workers in speaker identification learn how to cope with them.

### A summary of the state of the art

I would like to conclude by summarizing the state of the art in dealing with speaker differences.

Several speaker-independent and/or speaker-adaptive speech recognition systems have been successfully demonstrated, and a few commercial systems have been sold. In the next few years we should see commercial recognition systems with multi-speaker capability becoming increasingly common. There will, however, always be a proportion of the population who cannot use a particular system, either because of personal peculiarities in their voices or because they speak a form of the language too far removed from the forms on which the system was trained. Equally, I believe that it will remain true that the most reliable performance will be obtained by training a system on the voice of the person who is going to use it.

Speaker verification systems have been successfully demonstrated, but to my knowledge none have so far been sold. Their commercial appearance may be linked with the large-scale introduction of speaker-independent recognition systems allowing fully automatic remote access to information banks of one kind or another.

In my opinion, a useful level of automatic speaker identification has yet to be convincingly demonstrated on fully realistic, transmission-degraded material. We have to take into account the fact that, since the effectiveness of a speaker identification system would possibly be enhanced if its existence were unknown to the target group, publication of a major breakthrough in this area might be suppressed. Nevertheless, it is my guess that no major breakthrough has so far been made. It may be that automatic speaker identification may be eventually rendered feasible not by progress in the identification field itself but rather by the steady improvement in the quality of speech telecommunications systems.

### References

[1] MAIDMENT J.A. "Voice Fundamental Frequency Characteristics as Language Differentiators", *Speech and Hearing - Work in Progress,*, Dept. of Phonetics, University College, London, 1976.

[2] LAVER J.M.D.H. *Individual Features in Voice Quality*, PhD Thesis, Dept. of Linguistics and Phonetics, University of Edinburg, 1975.

[3] LADEFOGED P. & BROADBENT D.E. "Information conveyed by vowels", *J. Acoust. Soc. America*, Vol. 29, pp.98-104, 1957.

[4] VERBRUGGE R.R, STRANGE W., SHANKWEILER D.P. & EDMAN T.R. "What in $F_0$ enables a listener to map a talker's vowel space?", *J. Acoust. Soc. America*, Vol. 60 pp.198-212, 1976.

[5] KUHN G.M. "On the front cavity resonance and its possible role in speech perception", *J. Acoust Soc. Americaa*, Vol 58, pp.428-434 (1975).

[6] MONSON R.B. "Vowel normalization and ear canal resonance" *J. Acoust. Soc. America*, Vol. 66, p.S64 (1979).

[7] ALINAT P. "Phoneme recognition using a cochlear model", in *Spoken Language Generation and Understanding:* Proceedings of the NATO Advanced Study Institute held at Bonas, France, June 26-July 7, 1979, D. Reidel Publishing Co., Dordrecht, 1980.

[8] HUNT M.J. & MERMELSTEIN P. "Normalization of speech for automatic recognition", unpublished Final Report, Govt. of Canada DSS Contract No. 8SR79-00048, 1981.

[9] MOSHIER S.L., "Talker independent speech recognition in commercial environments", *J. Acoust. Soc. America*, Vol. 65 p.S132, 1979.

[10]SEARLE C.L., JACOBSON J.Z & RAYMENT S.G. "Stop consonant discrimination based on human audition", *J. Acoust. Soc. America*, Vol. 65 pp.799-809 (1979).

[11]LOWERRE B.T. *The Harpy Speech Recognition System*, PhD Thesis, Dept. of Computer Science, Carnegie-Mellon University, 1976

[12]KLATT D.H. "Speech perception: a model of acoustic phonetic analysis and lexical access", *Journal of Phonetics* Vol. 7 pp.279-312

[13]MOORE R.K., BEARDSLEY D., RUSSELL M.J. & TOMLINSON M.J. "Towards an integrated discriminative network for automatic speech recognition" *Proc. of the Institute of Acoustics*, London, (1982).

[14]RABINER L.R., LEVINSON S.E, ROSENBERG A.E. & WILPON G. "Speaker-independent recognition of isolated words using clustering techniques", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-27, pp.336-349, 1979.

[15]PETERSON G.E. & BARNEY H.L. "Control methods used in a study of the vowels" *J. Acoust. Soc. America*, Vol. 24, pp.175-184, 1952.

[16]LOWERRE B.T. "Dynamic speaker adaptation in the Harpy speech recognition system", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Hartford, CT, May 1977, pp. 788-790.

[17]FURUI S. "A training proceedure for isolated word recognition systems", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28 pp.129-136, 1980.

[18]HUNT M.J "Speaker adaptation for word-based speech recognition systems", *J. Acoust. Soc. America*, Vol. 69, pp.S41-42, 1981.

[19]FURUI S. "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-29, pp.254-272, 1981.

[20]BOGNER R.E. "On Talker Verification Via Orthogonal Parameters," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-29, pp. 1-12, Feb. 1981.

[21]NEY H., GIERLOFF R., & FREHSE R."An automatic system for verification of cooperative speakers via telephone", *Proc. Carnahan Conf. on Crime Countermeasures*, Lexington KY, pp.97-101, May 1981.

[22]DODDINGTON G.R. "Speaker verification", *RADC Technical Report* U1-963700-F, 1974.

[23]P.D. BRICKER, R. GNANADESIKAN, M.V. MATHEWS, S. FRUZANSKY, P.A. TUKEY, K.W. WACHTER & J.L. WARNER, "Statistical techniques for talker identification," *Bell Syst. Tech. J.*, Vol. 50, pp. 1427-1454, 1971

[24]SCHWARTZ R., ROUCOS S & BEROUTI M. "The application of probability density estimation to text-independent speaker identification", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, May 1982.

[25]PORITZ A.B. "Linear predictive hidden Markov models and the speech signal", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, May 1982.

[26]B.S. ATAL, "Automatic Recognition of Speakers from their Voices," in *IEEE Proceedings*, Vol 64, pp. 460-475, April 1976.

[27]HUNT M.J. "Further experiments in text-independent talker recognition over communications channels", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston MA, paper 12.12, April 1983.

# ASSESSMENT OF SPEECH SYSTEMS

Roger K. Moore

Royal Signals and Radar Establishment
St. Andrews Rd. Malvern
Worcestershire
U.K.

## SUMMARY

This lecture provides an overview of the key concepts relevant to the evaluation of speech transmission, speech synthesis and speech recognition systems. For speech transmission systems the concept of intelligibility testing is introduced, and techniques for both subjective and objective measurements are described briefly. It is then pointed out that very few standard tests or procedures exist for assessing either speech synthesis or speech recognition systems. So, after a quick look at the problems posed by speech synthesisers, the remainder of the lecture concentrates on automatic speech recognitior. The key issues in speech recogniser testing are discussed, and it is pointed out that the need to evaluate such systems has raised some very difficult questions, to which, as yet, there are few satisfactory answers; this area is currently a major research topic in the speech recognition community. It is shown how many interrelated factors affect the performance of a speech recognition machine, and that even interpreting and comparing experimental results can present some difficulties. Some tentative procedures are outlined and a scheme for estimating the relative difficulties of different vocabularies is described in detail. Finally, it is emphasised that evaluation techniques are crucial to the satisfactory deployment of automatic speech recognition equipment in real applications.

# A SURVEY OF EQUIPMENT AND RESEARCH

J.S. BRIDLE

JOINT SPEECH RESEARCH UNIT
PRINCESS ELIᴗABETH WAY
CHELTENHAM
ENGLAND GL52 5AJ

## 1.0 SUMMARY

Some terms for description of speech recognition systems are defined. A selection of real-time, commercially available speech recognition equipment is described, concentrating on the high-performance end of the market. Likely developments are indicated.

The lectures in the rest of this series have concentrated on a single approach to automatic speech recognition - that using whole-word templates. We explain current attempts to extend the capabilities of this approach, and also look at alternative approaches which are the subject of research in laboratories around the world.

## 2.0 INTRODUCTION

This chapter is intended to give the reader an idea of the sorts of differences that exist between current speech recognition equipments and research approaches.

Anyone considering using speech recognition in a real application such as in an aircraft cockpit should pay attention to at least the following three aspects. Firstly, assuming that the application needs a high-performance speech recogniser, ignore the low-cost end of the market; use one of the expensive, bulky, full-facility equipments in simulations of the real task, to find out what is required. Secondly, worry about special conditions in the task that could cause problems. Thirdly, consider how an ASR equipment can be acquired that will fit into the space available and interface with the other equipment.

This survey is a snapshot at a particular point in time. I shall try to indicate the way that things are moving, but you can assume that advances in micro-electronics will make available, in small, affordable packages, anything that can be done in real time now.

## 3.0 TERMS FOR DESCRIPTION OF SPEECH RECOGNITION MACHINES

The list below concentrates on aspects which concern the designer of systems which are to include a speech recognition equipment. All the systems below are Speaker Dependent: this means that they are meant to be set up separately for each speaker, by him saying all the words one or more times each.

### 3.1 Vocabulary size

Much confusion has been caused by use of this term to refer to very different aspects of a recogniser's capabilities. First we must distinguish between word and templates. Usually each template corresponds to one word (or a short phrase which can be used as if it were a single word) but one word may be represented by several templates.

The total number of templates that a system can hold depends only on the amount of memory in it, and its addressing range. The number of templates that the machine can compare with the input at any onᴗ time depends on its processing power. But the number of words that it can reliably distinguish between depends on its discrimination power, the similarity of the set of words, and the way they are said. The discrimination power is most interesting, but there are no established ways of defining it (although see Moore [1]).

In practice the number of words that a speaker-trained system can use is often limited by the need to acquire the template data. See 'training procedures' below.

### 3.2 Speaking style

The first commercial speech recognition equipments were isolated word recognisers. They required the user to pause between words for long enough to mark the end of each word. Because many words contain short gaps within them, the gaps between words had to be quite long, typically 300ms. The best systems impose a negligible extra delay for processing.

One attempt to improve things was called 'Quicktalk'. The user still had to leave a gap between words, but it could be smaller than the largest gap in a word.

Isolated-phrase connected word recognition systems use a pause (say 300ms) to signal end-of-phrase, and allow the words to be spoken without pauses. These systems usually have a maximum length of phrase, may have a maximum number of words in a phrase, and produce their answers after the final pause.

Continuous connected word recognisers dispense with the end-of-phrase detection, deal with pauses with the same mechanism as is used for word recognition, and can produce answers while the user is talking.

Isolated word mode is slow and unnatural. Any attempt to improve input speed is bound to run into the hard limit set by the end-of-word decision. It seems that the quicktalk method can increase input rate, but merely places the hard limit somewhere else. I have no data on the relative usefulness of the two connected word types. It is still necessary to learn a technique of speaking clearly, consistently, and not too fast, but connected word recognisers seem to fail more gracefully when pushed to their limits.

In published work on recognition performance [2,3,4] the best isolated word recognition performances have been by connected word recognisers.

### 3.3 Control of the recognition process

To get the best performance from a speech recogniser it is necessary to limit the number of different words considered at each point. In isolated word recognition the set of words to be considered can be controlled quite easily. Some systems allow the host system to specify the set of templates acceptable each time. Others have internal control, based on previous recognition decisions.

In connected word recognition it is not possible to decompose the recognition process into a sequence of decisions - the identity of each template in the best sequence can depend on the identity, and position in the input, of all the others. Some connected word recognisers can use a specification of the order in which the templates must appear (i.e. a grammar, or syntax) some can use just the identity of the templates that are acceptable in a string, and for some it is only possible to specify the number of words expected in a string. There is no uniformity yet in capabilities or terminology in this area, but there is no doubt that used wisely these techniques can greatly enhance performance in difficult conditions.

### 3.4 Information returned from the recogniser

The most basic information that a recogniser can return is the serial number of each template recognised. Many recognisers can also be set up to produce an arbitrary string of characters when a template is recognised, and this can be useful when driving an application program which was designed for keyboard input. (A recogniser with such capabilities is often described as voice input terminal).

A more specialised application program might be able to make good use of indications of reliability of recognition (scores) and alternative interpretations (with scores). The latter are more difficult to provide in a connected word recogniser than in an isolated word recogniser.

### 3.5 Training procedures

The process of acquiring template patterns generally goes by the rather confusing name of 'training'. Training methods are crucial to the success of template matching recognisers, because they have no other source of information about the set of words to be distinguished. (Training the user is also crucial to success, but is outside the scope of this lecture).

Some systems combine several training utterances of a word, to produce one, averaged, template. Other systems keep training tokens as separate templates, and recommend that two or three examples are used for the more difficult words (the digits usually). Systems that can make do with single example utterances are at an advantage in applications with many words that are quite distinct. Systems that can exploit the availability of many examples of each word are at an advantage in applications with a relatively small number of difficult words that must be recognised reliably. No current machine combines the advantages of both methods.

Some systems use a 'robust' training procedure that refuses to accept utterances which are very different from previous ones [5]. Connected word recognisers often use isolated words to form templates, but it seems that the use of examples from connected-word contexts can give better results, particulary if information about the variability of different parts of each word is captured and used.

### 3.6 Rejection of spurious inputs

In its most basic mode a recogniser will compare any sound with its current set of templates, and choose the most similar template. All recognisers have some facilities for rejecting sounds that are very different from all the templates, and sometimes also when no one template scores significantly higher than all the rest. There is usually an adjustable reject threshold, but the important thing is how well the recogniser will reject spurious noises and 'illegal' words while accepting valid words. There are no accepted tests for this aspect of performance.

### 3.7 Size, weight and cost

These factors depend on whether the recogniser is a chip set, a board, a terminal or a complete development system. The amount of support from the manufacturer is also very variable.

## 4.0 DESCRIPTIONS OF A SELECTION OF ASR EQUIPMENTS

Lea has produced a book [6] and a recent article [7] on selecting recognisers. The list below includes all connected word recognisers known to me, plus two high-performance isolated word recognisers which are of special interest.

### 4.1 NEC DP200 (Nippon Electric Co.Ltd., Japan)

A successor [8] of the first connected word recognition machine, the DP100 [9]. A self-contained isolated phrase connected word recogniser, with built-in tape storage. Vocabulary size 50 to 150 words in connected mode. Maximum duration of phrase 4s. Up to 5 connected words per phrase. Response 300ms from end of phrase. The set of words for each phrase can be controlled. One or two one-utterance templates per word. Can use connected word training.

### 4.2 MSDS SR128 (Marconi Space and Defence Systems, U.K.)

A self-contained isolated phrase connected word recogniser, with built-in tape storage. Template memory size 128 seconds. Maximum duration of phrase 10s. No limit to number of words in a phrase. Response 300ms from end of phrase. The set of words for each phrase can be controlled. One or two one-utterance templates per word. In use for flight trials in civil transport aircraft. Planned to fly in jet fighter in 1983.

### 4.3 Logica LOGOS (Logica Ltd., U.K.)

An equipment [10] designed for experiments on applications and human factors aspects. See lecture 'Inside a speech recognition machine' in this volume. Vocabulary store 100 to 2000 templates. Computation power for 25 to 200 templates. Continuous connected word recognition style. Response delay depends on ambiguity of input: typically one or two words delay. Recognition can be guided by word order syntax within phrases, with optional automatic switching to alternative syntaxes. Normally one or two one-utterance templates per word. Can use connected word training.

### 4.4 Verbex 3000 (Verbex Corp., USA)

A successor to a connected word recognition system which was implemented on the Verbex 1800 system. Using the 1800 system Verbex demonstrated very impressive connected digit recognition performance on recordings made for the US Postal Service [3]. The Verbex 3000 is aimed primarily at the industrial materials handling market, and very good noise tolerance is claimed. It uses a statistical word model, which generalises the idea of a template. Word models are built automatically using many examples of each word. Connected-word training material is normally used if the application calls for connected word input. Syntax control within phrases. Continuous-type algorithm, but output is produced 300ms after end of phrase.

### 4.5 Votan 1000,5000 (Votan, USA)

A fairly new isolated word recogniser [11], which uses dynamic programming. One or two utterances per word for training. Template memory size 500 seconds, maximum logical vocabulary size 256. Good performance in noise is claimed, and good results have been reported in simulated helicopter noise [12]. The 5000 also includes speech storage and replay.

### 4.6 Vecsys RMI88 (Vecsys France)

A commercial version of the MOISE experimental system from LIMSI [13]. A fairly new isolated word recogniser, which uses dynamic programming . Computation power for up to 125 templates. One or two utterances per word for training. Syntactic selection of vocabulary subsets. A related system has performed successfully in a jet fighter aircraft. A new version will offer connected word recognition [14].

## 5.0 A SURVEY OF RESEARCH IN AUTOMATIC SPEECH RECOGNITION

This section provides brief references to a selection of current research efforts on important topics.

### 5.1 Developments of whole-word template matching (WWTM)

Many groups are trying to extend WWTM methods, usually based on the dynamic time warp (DTW) application of dynamic programming (DP).

5.1.1 Connected words - There are several published algorithms which solve the mathematical problem of finding the best sequence of whole-word templates to match a given unknown speech pattern [15,16,17,18,19,20,21]. They differ in amount of storage and computatation needed, their ability to include in-phrase syntax control, the availability of alternative, sub-optimal 'explainations', and methods for 'pruning' the search process to reduce workload without significantly compromising the ability to find the correct answer. The most efficient connected-word algorithms need very little more computation than equivalent isolated-word algorithms.

5.1.2 Many talkers - The favorite method at present is to use several (e.g. 12) templates per word, chosen in an attempt to cover all pronunciations [22,23].

5.1.3 Operation with difficult signals - It is worth distinguishing between distorted speech, continuous background noise, and short duration high amplitude noises. Some manufacturers claim that their recognisers will work over the telephone system, which introduces non-linear and linear distortions. Many research laboratories are working with telephone speech. Continuous background noise is a problem in aircraft and elsewhere. The three main approaches are: subtraction of the noise waveform, using a second microphone and special filtering; subtraction of estimated noise power from the input spectrum; and allowing for the presence of the background noise when comparing template and input spectra.

5.1.4 Large vocabularies - There are many problems with the use of WWTM for vocabularies of more than a few dozen words [24]. Some groups have been most concerned with the training problem, and have resorted to the use of general-purpose units smaller than words [25]. Others have worried about the computational workload, and proposed initial sorting based on gross structure of the word. A more fundamental concern is for the discrimination power, and methods have been proposed for building into the 'templates' far more information about the variability of different -ts each word [26,27].

### 5.2 Largely automatic, statistics-based approaches

The IBM Continuous Speech Recognition team has been the main exponent of statistically based methods for many years [28]. Their goal is transcription of limited natural language (e.g. for business letter dictation). This is different from all other applications considered here, where the user is assumed to be prepared to use an artificial 'language' which is specific to the task. Some of the techniques have been described by Baker [29,30]. The Verbex 1800 and 3000 systems use word-based statistical models [31]. Recent work at Bell Labs has applied statistical methods successfully to speaker-independent isolated digit recognition [32].

## 5.3 Phonetics based approaches

Many attempts have been made to use speech knowledge in the design of speech recognition machines, but without much success. Reasons for this failure have included lack of adequate knowledge, difficulty in converting available knowledge into a form useful in speech recognition, and excessive complexity, leading to difficulty in testing and tuning the system. among groups currently attempting to produce useful recognisers based on phonetic principles are the National Physical Laboratory, UK, and Thomson-CSF, France.

A significant and well-equipped team at MIT is concentrating on identifying and quantifying specific knowledge about how the acoustic characteristics of speech sounds are affected by their context, and incorporating this knowledge into recognition procedures [33]. The goal is to eventually lift the barriers to speaker independence, large vocabularies and true continuous speech recognition.

Other sources of speech-specific knowledge are experimental psychology [34] and auditory neurophysiology [35].

## 6.0 CONCLUSIONS

The whole-word template matching method is likely to be the basis of practical speech recognisers for some years, and the cost and size of recognition components which perform as well as the best current systems can be expected to fall dramatically.

The best new systems will combine features of straightforward pattern matching and statistical modelling. For applications which need fast, reliable data entry, possibly in difficult conditions, but with a well-defined task and dedicated, trained users, such systems will be very suitable.

Other application areas will need recognisers that can exploit the regularities of speech sound structure to provide large vocabularies and the ability to adapt to the speech characteristics of a new speaker in a general way. It is likely that a combination of pattern matching, statistical modelling and phonetics will be needed.

## 7.0 REFERENCES

1. R.K.Moore, "Evaluating speech recognisers", IEEE Transactions on Acoustics, Speech and Signal Processing 25, No.2 (1977), pp.178-183.

2. Doddington, G.R. and Schalk, T.B., "Speech recognition: turning theory to practice", IEEE Spectrum, Vol.18, No.9, pp.26-32, September 1981.

3. Janet M.Baker, "How to achieve recognition: a tutorial/status report on automatic speech recognition", Speech Technology, Vol 1 No 1, pp 30-43, 1981.

4. H.F.Silverman, "Some general, user-oriented concepts for discrete utterance recognition (DUR) applications" IEEE 1982 International Conference on Acoustics, Speech and Signal Processing, Paris, May 1982.

5. Rabiner, L.R., Bergh, A. and Wilpon, J.G., "An embedded word training procedure for connected digit recognition", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Paris, 1982, pp.1621-1625.

6. W.A.Lea, "Selecting, designing and using speech recognisers", Speech Science Publications, Santa Barbara, Calif., 1983.

7. W.A.Lea, "Selecting the best speech recogniser for the job", Speech Technology, Vol.1 No.4, Jan/Feb 1983.

8. Tsuruta, S., et al., "DP-200 continuous speech recognition system", Reports of the 1981 Spring Meeting, Acoustical Society of Japan, pp.563-564, 1981.

9. Tsuruta, S., "DP-100 voice recognition system achieves high efficiency", JEE, pp.50-54, July 1978.

10. Peckham, J.B., Green, J.R.D., Canning, J.V. and Stephens, P., "A real-time hardware continuous speech recognition system", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Paris, 1982, pp.863-866.

11. Gill, S., "Low-cost development system recognises, stores, transmits and reproduces speech", Speech Technology, Vol.1, No.2, pp.79-80, 1982.

12. C.R.Coler, "Helicopter speech-command systems: recent noise test are encouraging", Speech Technology, Vol.1 No.3 (Oct.1982) pp.76-81.

13. J.-L.Gauvain, J.-S.Lienard and J.Mariani, "On the use of time compression for word-based recognition", IEEE 1983 International Conference on Acoustics, Speech and Signal Processing, April 1983.

14. J.L.Gauvain and J.J.Mariani, "A Method for connected word recognition and word spotting on microprocessor", IEEE 1982 International Conference on Acoustics, Speech and Signal Processing, Paris, May 1982.

15. Vintsyuk, T.K., "Element-wise recognition of continuous speech consisting of words of a given vocabulary", Kibernetika (Cybernetics), No.2, 1971.

16. Bridle, J.S., Brown, M.D. and Chamberlain, R.M., "A one-pass algorithm for connected word recognition", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Paris, 1982, pp.899-902.

17. Sakoe, H., "Two-level DP matching - a dynamic programming based pattern matching algorithm for connected word recognition", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol.ASSP-27, No.6, pp.588-595, December 1979.

18. Sakoe, H. and Watari, M., "Clockwise propagating DP-Matching algorithm for connected word recognition", Acoustical Society of Japan, pp.517-524, December 1981.

19. Myers, C.S. and Rabiner, L.R., "A level building dynamic time warping algorithm for connected word recognition", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol.ASSP-29, pp.284-297, April 1981.

20. Myers, C.S. and Levinson, S.E., "Connected word recognition using a syntax-directed dynamic programming temporal alignment procedure", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Atlanta, 1981, pp.956-959.

21. H.Ney, "Experiments in connected word recognition", IEEE 1983 International Conference on Acoustics, Speech and Signal Processing, April 1983.

22. Rabiner, L.R., Levinson, S.E., Rosenberg, A.E. and Wilpon, J.G., "Speaker-independent recognition of isolated words using clustering techniques", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol.ASSP-27, pp.336-349, August 1979.

23. Gupta, V. and Mermelstein, P., "Effects of speaker accent on the performance of a speaker-independent, isolated-word recogniser", J. Acoust. Soc. Am., 71, No.6, pp.1581-1587, 1982.

24. A.E.Rosenburg, L.R.Rabiner and J.G.Wilpon, "Speaker trained recognition of large vocabularies of isolated words", IEEE 1982 International Conference on Acoustics, Speech and Signal Processing, Paris, May 1982.

25. Hunt, M.J., Lennig, M. and Mermelstein, P., "Experiments in syllable-based recognition of continuous speech", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Denver, 1980, pp.880-883.

26. Moore, R.K., Russell, M.J. and Tomlinson, M.J., "The discriminitive network: a mechanism for focusing recognition in whole word pattern matching", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Boston, 1983.

27. Russell, M.J., Moore, R.K. and Tomlinson, M.J., "Some techniques for incorporating local timescale variability information into a dynamic time-warping algorithm for automatic speech recognition", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Boston, 1983.

28. L.R.Bahl, F.Jelinek and R.L.Mercer, "A Maximum likelihood approah to continuous speech recognition", IEEE Trans on Pattern Analysis and Machine Intelligence,

29. Baker, J.K., "The DRAGON system - an overview", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol.ASSP-23, No.1, pp.24-29, February 1975.

30. J.M.Baker and J.K.Baker, "Aspects of stochastic modelling for speech recognition", Speech Technology, Vol.1 No.4, Jan/Feb 1983.

31. P.F.Brown, C.-H.Lee and J.C.Spohrer, "Baysian adaptation in speech recognition", IEEE 1983 International Conference on Acoustics, Speech and Signal Processing, Boston, April 1983.

32. S.E.Levinson, L.R.Rabiner and M.M.Sondhi, "Speaker independent isolated digit recognition using hidden Markov models" IEEE 1983 International Conference on Acoustics, Speech and Signal Processing, Boston, April 1983.

33. D.W.Shipman and V.W.Zue, "Properties of large lexicons: Implications for advanced isolated word systems", IEEE 1983 International Conference on Acoustics, Speech and Signal Processing, April 1983.

34. E.Zwicker, E.Terhardt and E.Paulus, "Automatic speech recognition using psychoacoustic models", J.Acoust.Soc.Am., Vol.65 No.2 Feb.1979 pp.487-498.

35. R.F.Lyon, "A Computational model of filtering, detection and compression in the cochlea", IEEE 1982 International Conference on Acoustics, Speech and Signal Processing, Paris, May 1982.

# RECONNAISSANCE AUTOMATIQUE DE LA PAROLE DANS LES AVIONS D'ARMES

JEAN ROBERT COSTET

CROUZET S.A.
25 rue Jules Védrines
26027 VALENCE
FRANCE

## RESUME

Les études d'application des techniques de reconnaissance de la parole aux commandes de fonctions dans les avions d'armes viennent d'entrer dans une phase active. Les considérations développées dans le texte s'appuient sur des expérimentations, passées ou en cours, au simulateur ou en vol.

Les principaux thèmes développés seront les suivants :

- Les aspects techniques : la prise de son, le bruit et les problèmes posés par sa variabilité, l'importance de la phase d'apprentissage de la machine, les contraintes physiques.

- Les aspects opérationnels, en particulier les problèmes pratiques posés par l'intégration de la commande vocale dans les cabines d'avion.

- Les différentes expérimentations effectuées au simulateur et en vol et les enseignements qu'il est possible d'en tirer.

En conclusion, on tentera de dégager les perspectives qui s'offrent à ce type de technique dans le domaine aéronautique.

## 1 - PRESENTATION GENERALE

### 1-1 Le dialogue vocal dans un avion d'armes

Les travaux entrepris depuis quelques années sur l'application de la commande et de la synthèse vocale au dialogue pilote - système dans un avion d'armes ont suffisamment progressé pour qu'il soit envisageable de mettre en oeuvre ces techniques sur les avions de combat de la prochaine génération.

L'objectif poursuivi est de permettre au pilote de mieux se concentrer sur l'essentiel de sa mission dans un environnement opérationnel et technique de plus en plus complexe. Le but devra être atteint par une nouvelle conception du dialogue pilote avion, intégrant harmonieusement commandes vocales et manuelles ; ce dialogue devra à la fois être plus riche et plus souple, il devra permettre un contrôle facile des systèmes et des capteurs et offrir au pilote des possibilités de perception et d'analyse rapides et complètes des situations tactiques ainsi qu'une bonne capacité d'anticipation.

Cependant l'utilisation de la commande vocale se heurte encore à plusieurs difficultés, qui sont d'ordre technique (sûreté de fonctionnement) et fonctionnel (intégration dans une cabine) et qui devront être résolues dans les prochaines années.

### 1-2 Rappel des techniques utilisées

Les techniques de reconnaissance ayant fait jusqu'à présent l'objet d'expérimentation ou de tentatives d'expérimentations dans des milieux réels - et plus particulièrement en aéronautique - appartiennent toutes à la famille des reconnaissances monolocuteur dites "globales" ou encore "acoustiques" de mots. Leur caractéristique commune est de baser la reconnaissance de mots sur la comparaison entre la forme globale d'un mot qui vient d'être prononcé, et les formes similaires, dites "références" disponibles en mémoire. Ceci se traduit par la nécessité de créer ces références lors d'une phase particulière d'utilisation de la machine, dite phase d'apprentissage.

Jusqu'à présent, la Société CROUZET a mis en oeuvre au simulateur et en vol une technique de reconnaissance de mots isolés, dans un mode de reconnaissance dit "à micro commandé", qui utilise un bouton poussoir d'activation sur le manche ou sur la manette.

Les prochaines expérimentations utiliseront toutefois une technique de reconnaissance de mots connectés.

## 2 - LES ASPECTS TECHNIQUES IMPORTANTS

### 2-1 Les différentes sortes de bruit et les problèmes de prise de son

#### 2-1-1 Le bruit dans la cabine

Le bruit cabine a plusieurs origines :

- Le bruit du moteur, qui est transmis directement ou par l'intermédiaire de structures rigides. On peut encore distinguer les régimes à sec des régimes avec post combustion.

- Le bruit aérodynamique, qui est fonction de la vitesse et de l'altitude de l'avion ; à haute vitesse et basse altitude ($V_i > 450$ Kts) ce bruit peut devenir prépondérant sur les autres sources.

- Le bruit dû à la pressurisation et à la climatisation de la cabine, ce bruit est très variable et peut à l'occasion se révéler important.

- Le bruit des machines électriques : moteurs, ventilateurs, transformateurs.

D'une façon générale, le bruit d'un avion d'armes perçu dans la cabine est extrêmement variable d'un avion à l'autre : il existe des avions bruyants et des avions silencieux. Pour rendre leurs avions plus silencieux, pour améliorer le confort des pilotes mais surtout la qualité des transmissions radio, les constructeurs effectuent des analyses spectrales et temporelles fines du bruit dans la cabine.

Ces analyses détaillées sont indispensables à l'identification des sources de bruit et à leur traitement ; il n'est pas évident que ce soit le cas pour la reconnaissance de parole, pour trois raisons :

- le bruit régnant dans la cabine n'est pas celui perçu par le microphone de masque,

- la répartition spectrale du bruit est large et il serait illusoire d'espérer agir localement,

- l'énergie du bruit et sa répartition spectrale sont suffisamment variables (parfois au cours d'un même vol) pour qu'il soit plus profitable de s'intéresser à son enveloppe et à son mode de variation qu'à sa composition fine, sauf si on désire agir sur les sources de bruit elles-mêmes.

A titre d'exemple, le niveau de bruit le plus élevé mesuré au point fixe dans la cabine de l'avion qui supportait notre expérimentation (Mirage III) était de 106 dBA. Aucune mesure n'a été faite en vol, mais le bruit croissait rapidement avec la vitesse et dépendait de l'altitude (maximum pour 540 Kts, 10000 ft) ; le bruit occupe toute la bande spectrale, avec un maximum entre 500 et 2500 Hz.

#### 2-1-2 Le bruit dans le masque à oxygène

Ce bruit est très différent du bruit régnant dans la cabine, qui se trouve amorti par l'enveloppe de caoutchouc du masque ; ce dernier reste pourtant présent.

La source de bruit la plus importante est liée à la respiration du pilote.

- l'effet sonore produit par la circulation gazeuse dans l'espace confiné du masque et des tuyaux et à travers les clapets est important, il peut de plus être différent en fonction des types de masque,

- le microphone est souvent fixé dans le masque en face de la bouche et du nez du pilote et reçoit directement le souffle d'expiration de ce dernier.

Le bruit de respiration est particulièrement gênant car :

- il apparaît brutalement, avec un niveau d'énergie élevé

- il occupe lui aussi la bande spectrale de parole et plus particulièrement les hautes fréquences

- il peut aléatoirement se superposer à un mot ou le prolonger.

Si une oreille humaine distingue facilement ce type de bruit de la parole auquel il se mélange, il n'en va pas de même pour une machine de reconnaissance automatique de parole, qui ne sait généralement pas faire la distinction entre les deux types de signaux.

### 2-1-3 Les microphones et les problèmes de prise de son

. Les capsules microphoniques elles-mêmes sont en général satisfaisantes sur le plan de la dynamique et de la bande passante (supérieure à 5000 Hz), et leur courbe de réponse est souvent bien adaptée aux cas d'utilisations : par exemple les microphones de masque présentent une accentuation de la courbe de réponse dans les basses fréquences pour pallier un affaiblissement dû au masque lui-même.

. Les micros différentiels permettent d'abaisser la sensibilité aux sources de bruit étendues ou lointaines.

. Le signal électrique délivré par le micro doit être un signal de fort niveau, de manière à être le moins possible sensible aux perturbations électriques.

Les caractéristiques nominales des microphones modernes sont donc en général correctes, mais par contre :

- il peut y avoir des dispersions sur les micros utilisés et leur montage dans le masque

- le signal du microphone chemine souvent de façon compliquée par des systèmes d'interconnexions, les téléphones de bord, les contrôles de gains automatiques, etc..., qui peuvent le dénaturer et perturber la reconnaissance. On aura donc intérêt à établir une liaison filaire directe et bien protégée entre le micro et le calculateur de reconnaissance de parole.

. La prise de son ne dépend pas uniquement des caractéristiques du micro, mais également du masque et de son montage sur le casque. Les positions géométriques relatives de la capsule et de la bouche sont à considérer. Les micros actuellement utilisés en vol par exemple, sont placés directement devant et la bouche et le nez du pilote et sont donc beaucoup plus sensibles aux souffles produits par l'expiration ou l'inspiration que s'ils étaient placés latéralement.

## 2-2 Le mélange du bruit et de la parole et son traitement

### 2-2-1 L'intensité du bruit

On a pu obtenir un fonctionnement satisfaisant de la reconnaissance automatique de parole dans des environnements très bruités (atteignant 109 dBA, ou 115 dBC).

La proximité du micro et sa directivité fait, même dans ce cas, qu'il est possible de travailler avec des rapports signal/ bruit utilisables (environ 10 dB) dans la mesure où,au moins sur de courtes périodes (durée de la prononciation d'un mot), le bruit peut être considéré comme stationnaire.

L'intensité du bruit n'est donc pas forcément le seul critère à prendre en compte dans le problème de la reconnaissance automatique de parole dans le bruit.

### 2-2-2 La variabilité du bruit

Une grande partie du problème réside dans la variabilité du bruit et de ses modes d'apparition.

Il existe des variations lentes du bruit, dues par exemple aux variations de vitesse et de régime aérodynamique de l'avion ; mais il peut également apparaître de façon brusque dans plusieurs circonstances :

- Glissement sous facteur de charge d'un masque mal ajusté. Le bruit régnant dans la cabine peut alors être perçu de façon beaucoup plus forte par le micro.

- Le bruit de respiration. Dans l'espace confiné du masque à oxygène et des dispositifs associés, l'écoulement gazeux est canalisé à l'inspiration et à l'expiration et émet un bruit caractéristique, qui prolonge ou se superpose à la parole. Dans certains cas, lorsque le masque n'est pas correctement appliqué sur le visage, la dépression régnant dans la cabine peut conduire à un débit permanent et bruyant.

Dans les cas que nous venons de citer, ce n'est pas tant l'intensité du bruit qui est gênante que sa variabilité. Même faible, il peut pour la machine de reconnaissance apparaître comme un constituant du signal utile de parole ; il est en général d'autant plus difficile de distinguer du bruit dans un signal de parole que ce bruit ne posséde pas de caractéristiques stables permettant d'effectuer une prédiction.

De plus il peut se faire que même à l'oreille le bruit se distingue mal de certains sons vocaux comme les fricatives ou les sifflantes par exemple.

### 2-2-3 Les solutions possibles

Le mélange du bruit et de la parole pose deux types de problèmes à une méthode de reconnaissance de mots isolés :

- la détection de début et de fin de mot
- la reconnaissance proprement dite une fois sa détection faite.

. La détection de début et de fin de mot

En général, et surtout s'il s'agit de bruit non stationnaire, l'utilisation d'un simple seuil basé sur un niveau d'énergie ne donnera pas de bons résultats. Il faudra plutot rechercher un critère basé sur la différence de nature entre la parole et le bruit. Bien souvent, même dans le cas de bruits non stationnaires, la variabilité spectrale du bruit est inférieure à celle de la parole. Des critères de déclenchement basés sur cette propriété ont donné de bons résultats. Il reste toutefois difficile de déterminer la fin de certains mots, surtout des mots courts, dont la prononciation se termine par une forte expiration qui se mélange à eùx et les prolonge. C'est le cas en français pour des mots courts se terminant par une sifflante ou une fricative, exemple : six, neuf...

On peut cependant noter que l'entrainement du locuteur et le fait qu'il soit averti de ces problèmes est un facteur favorable à la maîtrise de l'élocution.

Pour les cas vraiment difficiles qui restent, la tendance serait plutot de rechercher des algorithmes qui ne nécessitent pas de détection précise de début et de fin de mot, c'est à dire des algorithmes de type reconnaissance de mots enchainés.

. La reconnaissance proprement dite

Une fois la détection de début et de fin de mot résolue, on peut envisager deux cas de signaux de parole bruités : le cas ou le bruit peut être considéré comme stationnaire, et celui où il ne l'est pas.

Dans le premier cas, le bruit peut être considéré comme stationnaire si ses caractéristiques ne varient pas pendant toute la prononciation du mot, ceci s'applique aux bruits d'origine aérodynamique, qui évolue lentement. De bons résultats ont été obtenus sans traitement particulier pour des rapports signaux sur bruit suffisamment grands (cf 2-2-1) Des techniques simples de soustraction de spectre de bruit ont également été expérimentées avec succès.

A noter un point important : une ambiance sonore élevée entraine une déformation de la voix du locuteur si celui ci la perçoit de façon trop importante (si la protection du casque est insuffisante);la tendance naturelle du locuteur est en effet de rétablir un rapport signal/bruit à peu près constant. Ces déformations sont vites sensibles et perturbent la comparaison dynamique. Deux précautions sont à prendre :

1) Vérifier que la protection du locuteur est à la fois suffisante et constante.

2) Faire l'apprentissage dans des conditions représentatives d'un cas moyen. Ultérieurement l'apprentissage adaptatif offrira peut être des solutions plus souples.

Les problèmes de la reconnaissance de la voix déformée en ambiance bruitée sont fondamentaux et liés à la représentativité des paramètres extraits lors de l'analyse du signal et à la qualité de la comparaison dynamique.

Dans le deuxième cas, la reconnaissance est difficile et les progrès à faire sont d'ordre fondamental, ils concernent la discrimination du bruit et de la parole mélangés, et l'amélioration de la comparaison dynamique. Si la parole et le bruit ne sont pas superposés, mais temporellement juxtaposés (bruit du souffle), il est possible d'améliorer les critères de sélection des paramètres délivrés par l'analyse du signal.

Dans les essais en vol actuellement en cours les problèmes de reconnaissance dus au bruit résultaient dans une proportion importante (30 %) d'erreurs de détection de début et de fin de mot.

## 2-3 L'apprentissage des références

### 2-3-1 L'importance de l'apprentissage

Dans les méthodes de reconnaissance globales, surtout en mots isolés, l'importance de l'apprentissage est fondamentale. En effet le niveau des performances atteint par la suite dépend de la bonne représentativité des formes acoustiques acquises lors de l'apprentissage, c'est à dire de leur ressemblance acoustique avec les mots tels qu'ils sont effectivement prononcés en vol.

Cette exigence de similitude n'est pas très grande tant qu'on fait de la reconaissance de parole dans des conditions de laboratoire, mais apparaît clairement dès qu'on aborde des milieux plus réalistes, tels que le simulateur ou le vol sur avion d'armes.

Rappelons que les techniques de reconnaissance de parole disponibles aujourd'hui ont des procédures d'apprentissage variées qui diffèrent, en particulier sur le nombre de passes d'apprentissage nécessaires (10 constituent un maximum). La méthode utilisée par Crouzet ne nécessite qu'une seule passe d'apprentissage.

L'expérience nous a montré que :

. La phase d'apprentissage présentait un certain caractère aléatoire (peut être lié au fait qu'il se fait en une passe unique)

. Plusieurs facteurs influaient sur la bonne représentativité des références issues de la phase d'apprentissage. Ces facteurs touchent l'environnement et sont d'ordre acoustique et ergonomique.

Le facteur acoustique : il nous paraît nécessaire que l'apprentissage soit effectué dans une ambiance bien représentative du niveau de bruit moyen de l'application, ce qui signifie qu'il s'agit de l'ambiance réelle ; dans notre cas, l'apprentissage est fait dans le cockpit, moteur en route, ou mieux, en vol.

Les facteurs ergonomiques :

- l'énoncé d'une liste de mots devient vite fastidieuse, et le naturel de l'élocution s'en ressent surtout si la liste est longue. De ce point de vue, la possibilité d'une unique passe d'apprentissage est favorable.

- L'apprentissage doit être effectué avec le même matériel et dans la même ambiance que ceux de l'utilisation ultérieure.

A titre d'exemple et pour illustrer l'importance de ces différents facteurs, on peut citer le fait que le pourcentage de mots reconnus en vol a été amélioré de façon importante (+ 5 %) à partir du moment où l'apprentissage ne s'est plus fait au sol, moteur tournant, mais directement en vol.

2-3-2 L'apprentissage évolutif

Pour plusieurs raisons, il serait souhaitable de faire évoluer l'ensemble initial des références issues de l'apprentissage.

- Certaines de ces références peuvent à l'usage se révéler médiocres.

- L'entrainement progressif du locuteur le conduit à parler d'une façon souvent différente de celle qu'il avait lors de l'apprentissage.

- En utilisation, les conditions d'environnement - surtout le bruit - peuvent évoluer par rapport à celles de l'apprentissage.

Il peut donc s'avérer nécessaire, pour obtenir un niveau de performances élevé :

- soit de retoucher le jeu de références initial

- soit de recréer un ensemble de références à partir de mots prononcés en vol lors de phases de reconnaissance, ce qui nécessite de les conserver en mémoire

- soit d'utiliser des procédures automatiques pour faire évoluer les références et les adapter en permanence aux évolutions de l'environnement ou de la voix du pilote. De telles procédures sont actuellement à l'étude.

## 2-4 Les contraintes physiques et physiologiques :

### 2-4-1 Les accélérations

Les effets des accélérations sur la reconnaissance de parole ont été étudiés dans un domaine présumé utile allant jusqu'à environ 4,5 g. Cette limite a été choisie pour deux raisons :

- il est difficile de soutenir sur Mirage III des virages continus à des inclinaisons plus fortes, or l'étude nécessite qu'un nombre suffisant de mots soient prononcés sous des facteurs de charge relativement constants.

- Les facteurs de charge plus élevés correspondent à des manoeuvres particulières, généralement de combat.

Dans les phases de combat, toutes les commandes dont le pilote a besoin se trouvent regroupées sous ses doigts, sur la poignée de manche et la manette des gaz. Seul l'appel de quelques paramètres particuliers serait justifiable  de hauts facteurs de charge.

Schématiquement, les problèmes résultants d'accélérations verticales importantes (mises en virages) sont de deux types :

- Les problèmes de respiration.

- Les problèmes de bruit.

Les accélérations affectent le cycle respiratoire ; le pilote ne respire plus normalement, mais inspire avant de parler, se contracte et parle de façon souvent rapide et saccadée. Les plosives initiales sont plus marquées ; le souffle de l'expiration soutenue peut se superposer de façon plus importante à la fin des mots. Cependant, jusqu'à 4 g ces effets sont encore faibles et peu décelables, même à l'oreille, et il semble possible d'aller plus loin.

L'augmentation du bruit est par contre assez sensible ; la cause principale en est le glissement du masque (voir § 2-2-2), les effets secondaires étant dus aux modifications du bruit d'origine aérodynamique dans la cabine (augmentation de l'incidence).

Actuellement, nous obtenons sous facteur de charge allant jusqu'à 4 g le taux de reconnaissance de l'ordre de 5 % inférieur à ceux obtenus en palier.

Il faut remarquer que la reconnaissance sous facteur de charge peut être influencée par une forte dispersion du comportement des pilotes dans ces conditions.

2-4-2 Les autres facteurs

Nous n'avons pas analysé sérieusement d'autres facteurs d'influence que les accélérations hormis l'altitude et le Mach, qui ne semblent pas se traduire par des effets appréciables.

Un facteur influent pourrait être la peur, mais il n'y a aucune raison a priori de penser que la peur affecterait davantage l'élocution que la faculté de raisonner, la motricité, l'habileté manuelle , etc ... Si l'utilisation de la commande vocale permet de simplifier, de rationnaliser et de clarifier les procédures de commande, de permettre par un dialogue plus riche une meilleure analyse et une meilleure anticipation des situations tactiques, pourquoi ne pas penser au contraire, qu'elle apporterait une aide supplémentaire ?

## 3 - LES ASPECTS OPERATIONNELS

### 3-1 Le gain

Les différents aspects du gain escompté ont été évoqués en introduction : rappelons les principaux :

. Meilleure concentration du pilote sur sa tâche à court terme, particulièrement quand le champ visuel est fréquemment sollicité par l'extérieur (combat, attaque au sol, approche, patrouille serrée...) ou par les visualisations électroniques.

. Augmentation de la sécurité dans le cas de pilotage difficile, où la vision de l'extérieur est essentielle (vol basse altitude, phases d'attaque, patrouille).

. Augmentation du degré d'interactivité du pilote avec les systèmes embarqués.

. Gain de temps et plus grande facilité dans l'exécution des procédures de commande.

. Augmentation des possibilités d'aménagement de la cabine, notamment dans les cas où on envisage l'utilisation de sièges très inclinés.

. Diminution du nombre de commandes manuelles.

Il ne sera toutefois possible d'obtenir ces gains et de tirer pleinement partie de la commande vocale que lorsqu'une intégration correcte dans la cabine pourra être réalisée.

### 3-2 L'intégration de la commande vocale dans une cabine d'avion-d'armes

La commande vocale s'adresse essentiellement aux avions dont la cabine reste à concevoir, il est illusoire d'en espérer un gain important dans une cabine équipée trop traditionnellement. Plusieurs considérations interviennent :

#### 3-2-1 Le parallélisme des procédures vocales et manuelles

La commande vocale n'étant pas absolument fiable, elle ne peut être le moyen unique de commander une fonction ; toute commande vocale doit donc pouvoir également être réalisée manuellement. Les commandes manuelles étant ici utilisées moins fréquemment, il est possible d'augmenter leur centralisation et leur degré de multiplexage et de diminuer le nombre de postes de commandes spécialisés ; les deux moyens de commande sont alors sur un même pied d'égalité, et leurs utilisations se ressemblent. Dans ce contexte, les commandes manuelles ne sont pas des secours des commandes vocales, mais chaque procédure de commande peut être réalisée à la voix ou manuellement. De plus, une séquence de commandes peut être entamée à la voix et poursuivie manuellement, ou l'inverse.

#### 3-2-2 La compatibilité mécanique

Les organes mécaniques des commandes qui peuvent être réalisées manuellement ou à la voix ne peuvent être quelconques ; les poussoirs à enfoncement possédant deux positions mécaniques stables, les basculeurs et les rotacteurs sont exclus ; les dispositifs utilisables ne doivent pas avoir plusieurs positions mécaniques stables ; on utilisera des touches à appui fugitif et éclairement, des commandes incrémentales, etc...

#### 3-2-3 La souplesse de dialogue

Les expérimentations au simulateur et en vol nous ont montré que les procédures de commande vocale devaient soigneusement être mises au point pour éviter les difficultés de fonctionnement. Nous en avons retiré plusieurs enseignements, par exemple :

- La reconnaissance de mots enchainés est très souhaitable. Elle permet :

. de se rapprocher de l'élocution naturelle

. de réaliser des commandes complexes (les commandes avec introduction de données numériques par exemple, sont longues et difficiles à réaliser avec des mots isolés)

. d'augmenter les possibilités de la commande vocale.

- L'obtention d'un bon niveau de sécurité nécessite l'emploi d'un bouton d'ouverture du micro.

- Dans l'utilisation combinée du bouton d'ouverture du micro et de la phase de commande, les hésitations, les silences, et les retards du locuteur doivent être possibles sans problème.

## 3-3 Les contraintes liées à l'apprentissage

### 3-3-1 L'apprentissage à bord

L'obtention de bonnes performances nous paraît aujourd'hui nécessiter que l'apprentissage se fasse à bord, et si possible, en vol. Il y a là une contrainte, mais il est probable que cet apprentissage se fera lors de vols d'entrainement ou de transformation. De plus l'apprentissage effectué par le pilote sera valable pour tous les avions du même type sur lesquels il sera susceptible de voler.

### 3-3-2 Le support des références

Les références issues de l'apprentissage sont attachées au pilote et doivent pouvoir être stockées sur un support quelconque : module mémoire électroniques, badge magnétique, cassette, ..., qui lui est personnel, qui doit pouvoir être transporté facilement et inséré avant chaque vol dans un lecteur situé sur l'avion ; il doit être possible d'en obtenir rapidement une copie en cas de perte.

La gestion d'un tel élément entraine des contraintes qui ne sont pas négligeables, mais les moyens qu'elle suppose ne sont pas considérables devant ceux que nécessiteront demain la préparation d'une mission.

## 3-4 Les fonctions

Les fonctions présentes sur avion d'armes dans lesquelles il sera intéressant d'inclure des procédures vocales sont très nombreuses, mais rien de précis ni de définitif ne peut être avancé aujourd'hui.

Citons simplement à titre d'exemple des fonctions qui peuvent être concernées :

- communication
- identification
- navigation
- préparation armement
- gestion des capteurs (en particulier dans les modes d'attaque Air-Air et Air-Sol)
- gestion de visualisation
- changement de modes
- consignes de pilotage
- interrogation de paramètres
- etc...

## 4 - L'EXPERIENCE EMBARQUEE "EVA"

## 4-1 Objectifs expérimentaux

Les essais en vol d'un système de reconnaissance de mots isolés venaient après une étude en simulateur, où l'aspect opérationnel avait été abordé.

L'expérimentation d'EVA (Equipement Vocal des Avions) avait pour but la mise au point et la validation dans l'environnement réel de la technique de reconnaissance de mots isolés.

## 4-2 Description d'EVA

### 4-2-1 Matériel embarqué

Le matériel embarqué en pointe avant du Mirage III R n° 306 du Centre d'Essais en Vol de Brétigny se compose de :

- Un boîtier de dialogue vocal (1/2 ATR court) ; il réalise la reconnaissance de mots isolés et la synthèse des messages et des pannes.
- Un lecteur/enregistreur de cassettes numériques Qantex. Ces cassettes contiennent :
  - . les références du pilote issues de l'apprentissage
  - . la forme numérique de tous les mots prononcés en vol (sonagrammes)
  - . 3 paramètres de vol (altitude cabine, le facteur de charge, le roulis)
- L'alimentation du Qantex
- Un magnétophone.

En cabine, un boîtier de commande et de visualisation situé en haut de la planche de bord canalise le déroulement de toutes les procédures de dialogue vocal et assure leur contrôle. Il permet :

- La mise sous tension d'EVA
- l'arrêt de la synthèse
- le choix du mode, apprentissage ou reconnaissance
- la visualisation sur écran à cristaux liquides (2 lignes de 20 caractères) des résultats de reconnaissance et d'informations diverses liées au dialogue.

### 4-2-2 Banc sol

Un banc spécifique permet d'effectuer sur le site :

- La prépa.ation des cassettes
- La première analyse et l'édition d'un vol d'essai.

### 4-3 Les fonctions concernées sur l'avion

La mise au point de la reconnaissance de mots isolés dans un environnement réel nécessite la présence de boucles de commande réalistes ; elles sont indispensables à la motivation du locuteur.

Les fonctions concernées sont les suivantes :

Appel de paramètres : le pilote peut demander vocalement la valeur de certains paramètres, cette valeur est alors simultanément affichée sur le boîtier de commande et visualisation et synthétisée dans le casque.

- le mach
- l'altitude
- l'incidence et le roulis
- la vitesse propre, et la vitesse indiquée
- la distance et le relèvement par rapport à une balise sélectée
- le facteur de charge
- le carburant restant.

Sélection de fréquences radio UHF : Les fréquences radio peuvent être :

- appelées par un nom de code
- composées (chiffre par chiffre)
- mises en service.

Autocommande (stabilisateur de trajectoire) : Il est possible :

- d'embrayer et de débrayer le stabilisateur de trajectoire
- d'embrayer et de débrayer la tenue d'altitude.

Synthèse de pannes : 10 pannes font l'objet d'une synthèse, chaque message de panne peut être activé individuellement par un interrupteur situé sur un panneau de commande spécifique.

Synthèse de changement d'état :

- Sortie et verrouillage du train
- Transfert de réservoir de carburant
- Débrayage de l'autocommande.

Le vocabulaire nécessaire à l'exécution de ces fonctions est composé de 22 mots, plus des 10 chiffres.

## 4-4 Déroulement des essais

### 4-4-1 Coordination essais - études

EVA possède un caractère très expérimental et les essais en vol interagissent profondément avec les études en laboratoire.

La mise au point de la reconnaissance de mots isolés dans l'environnement cabine utilise deux niveaux d'analyse des résultats :

- sur le site à l'aide du banc sol (premier diagnostic)
- dans les laboratoires de la société, à Valence.

Les données acquises en vol y sont intégralement conservées ; elles sont utilisées en simulation, pour mettre au point une technique, en étudier les modifications, et améliorer les performances.

### 4-4-2 Déroulement des essais

Les essais se sont déroulés sur environ 40 vols, à deux pilotes.

On peut approximativement y distinguer les étapes suivantes :

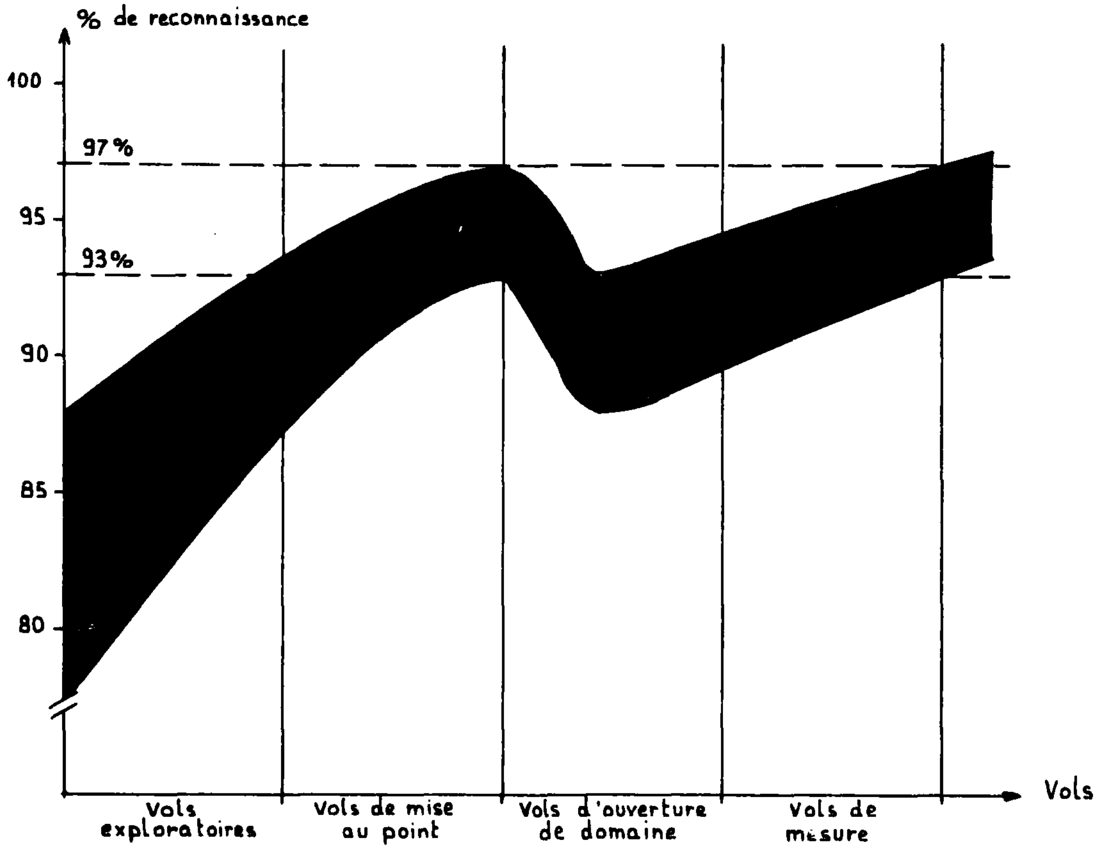
- Vols exploratoires avec vocabulaire et domaine de vol restreints.
- Vols de mise au point (mise au point de la méthode, du dialogue fonctionnel, et du vocabulaire).
- Vols d'ouverture du domaine de vol.
- Vols de mesure.

Compte tenu des évolutions continuelles de l'expérience au cours de la mise au point, pour la plupart des vols l'apprentissage s'effectuait dans l'avion, soit avant le vol, soit au tout début du vol. Seuls les vols de mesure se sont effectués sans apprentissage, en utilisant des références stockées sur une cassette.

## 4-5 Résultats

Les résultats obtenus ont été jugés de façon satisfaisante par les pilotes. Plusieurs problèmes, notamment d'acquisition ont dû être résolus durant la première phase des essais et lors de l'ouverture du domaine de vol.

On peut représenter l'évolution des performances obtenue par la figure ci-dessous :



Cette figure montre l'enveloppe moyenne des résultats d'essais en vol. L'allure générale ascendante est typique de l'amélioration au cours du temps, due à l'accumulation de l'expérience et à une mise au point continuelle.

Les dispersions à l'intérieur de l'enveloppe restent encore à analyser finement. Les facteurs humains y sont sans doute pour une part importante.

On notera la brusque détérioration des résultats au moment de l'ouverture du domaine de vol vers les facteurs de charge élevés et les grandes vitesses.

L'ascendance de la courbe nous permet d'espérer de nouveaux gains sur les résultats.

## 5 - CONCLUSION

Les perspectives ouvertes par le comportement de la reconnaissance de parole au simulateur et en vol sont très intéressantes et permettent d'entrevoir les premières applications en vol. Il est cependant nécessaire de travailler activement dans au moins deux directions :

- La maîtrise de la sécurité du dialogue.

. Les résultats déjà obtenus en vol sont très encourageants et montrent qu'il est probablement possible d'aller plus loin et d'obtenir des scores plus élevés encore.

. Cette maîtrise ne pourra être obtenue qu'à travers des expérimentations en vol intensives, et avec le plus de pilotes possibles, de manière :

- à mieux comprendre la diversité des problèmes qui se posent, même s'ils sont apparemment de faible importance,

- à cerner davantage la variabilité du signal de parole et du bruit

- à permettre une mise au point continue des techniques de reconnaissance.

En effet la reconnaissance automatique de parole est un des domaines où il est le plus difficile de reproduire et d'étudier en laboratoire les conditions d'utilisation dans un milieu réel.

- L'intégration des commandes vocales dans les cabines d'avions.

Cet aspect nécessite une compréhension approfondie de la nature et des possibilités du dialogue vocal afin de déterminer son rôle dans un cockpit. Le gain opérationnel doit être nettement dégagé ; ceci passe vraisemblablement par une conception nouvelle et globale des commandes.

Il est probable que dans un premier temps, la commande vocale se voie attribuer un rôle limité à quelques fonctions (Radio Communications et Radio Navigation par exemple ) puis que ce rôle soit ultérieurement étendu à d'autres fonctions, dans le cadre de cabines d'avions radicalement nouvelles.

Si ces travaux se poursuivent à un rythme satisfaisant, on devrait voir déboucher les premières applications opérationnelles à la fin des années 80.

# INSIDE A SPEECH RECOGNITION MACHINE

J.S. BRIDLE

JOINT SPEECH RESEARCH UNIT
PRINCESS ELIZABETH WAY
CHELTENHAM
ENGLAND GL52 5AJ

## 1.0 SUMMARY

The aim of this lecture is to illustrate some of the speech recognition techniques that have been presented so far, by concentrating on a particular speech recognition system that the author knows well. This system, known as Logos (from the Greek for "word"), is designed as a flexible, high-performance, experimental machine for research on recognition methods and applications aspects. However, it can serve as a point of reference when considering more practical machines, both current and future.

We first present the algorithms for connected word recognition on which the system is based. We then consider some of the practical matters that can be important in implementing such algorithms in computer programs and special-purpose equipment. Finally, we give an overview of the hardware system architecture, pointing out ways that the properties of the algorithms have influenced the design.

## 2.0 SYMBOLS USED

| | |
|---|---|
| M | Number of frames in the input pattern. |
| R | Number of templates in use. |
| N(r) | Number of frames in the r'th template. |
| dist(t,i,j) | Spectrum distance between the i'th frame of the t'th template and the j'th input frame. |
| C(t,i,j) | Sum of spectrum distances for the best explanation of the first j input frames, leading to the i'th frame of the t'th template. |
| T(j) | Identity of the last template in the best explanation of the first j input frames. |
| F(j) | Input frame at the end of the template preceding T(j) in best explanation of the first j input frames. |
| L(t,i,j) | Word link - number of input frame corresponding to the end of the template preceding the t'th template. |

## 3.0 THE CONNECTED WORD RECOGNITION ALGORITHMS

The obvious way to apply whole word template matching techniques to connected word recognition is to segment the input into words, then determine the identity of each word. Because segmentation is so difficult, the alternative approach we favour is to extend whole word pattern matching to deal with connected words in a natural way, without the need for a prior segmentation into words. We define the best word sequence for a given input utterance as the one for which the corresponding templates, when joined together end-to-end, make a "composite template" which matches the input pattern best (i.e. better than any other template sequence). This is illustrated in Fig. 1, where the sound pattern of the connected sequence "one-six-three-five-two" is displayed above a concatenation of the corresponding isolated digit patterns. Connected word recognition using whole word pattern matching techniques will be possible if the best word sequence corresponds to the actual words spoken often enough to be useful and if there also exists an efficient algorithm for finding the best word sequence.

We describe an efficient one-pass dynamic programming algorithm to find the sequence of templates which best matches the whole of the unknown input pattern [1]. We find that this approach works quite well, considering its naivety.

In the following description of the connected word recognition algorithm, it is assumed that the input and the template patterns are represented as sequences of units which are called "frames" (from vocoder terminology). In most of our work, each of these frames has been a 19-point spectrum cross-section. Each frame in Fig. 1 corresponds to 20ms of speech signal.

The main requirement for the matching algorithm is that any frame of the input can be compared with any frame of any template to give a measure of "distance" or "dissimilarity" between the two frames. Some distances for a simple example are plotted in Fig. 2. The input word sequence "one-three-two" is plotted with time axis horizontal and templates for "one", "two" and three" are plotted with time axis vertical. A distance of zero between two frames (i.e. identical data) is displayed as the largest size of black dot, and a range of other distances is displayed by using smaller dots. Outside this range the display is white.

## 3.1 Dynamic Programming Scoring for Connected Word Recognition

In isolated word recognition a "time registration path" maps the timescale of the input on to the timescale of a single template. In connected word recognition, a time registration path maps the timescale of the input on to the timescale of a sequence of templates. This is illustrated in Fig. 3 for a particular sequence of templates. This diagram can be re-drawn as in Fig. 4, where it can be seen that any other template sequence could also be accommodated. In Fig. 4 any valid path starts at one of the points A,B,C (the start of the input and the start of one of the templates) and ends at one of the points X,Y,Z (the end of the input and the end of one of the templates).

Within templates the time registration path can repeat or skip template frames, but transitions between templates are always to the start of one template from the ends of those templates which are permitted to precede it.

Scoring an arbitrary connected-template path is very similar to the isolated-word case: a simple sum is formed of all the between-frame distances along the path. The best explanation of the input corresponds to the best-scoring path.

The connected word recognition score, C(t,i,j), is defined as the sum of the distances for the best way of matching the first j input frames with any permissible sequence of templates followed by the first i frames of the t'th template (Fig. 5). Thus although the score, C(t,i,j), is still only a function of position (j) in the input and of position (i) in a template (t), as for isolated word recognition, it also depends on the other templates and the way they might explain previous parts of the input.

Within each template the same basic operation is performed as in isolated word recognition:

$$C(t,i,j) = \underset{a=0,1,2}{\text{Minimum}} \; C(t,i-a,j-1) + \text{dist}(t,i,j) \quad .\;.\;.\;.\; (1)$$

where dist(t,i,j) is the spectrum distance between the i'th frame of the t'th template and the j'th input frame.

At the start of each template, the ends of the preceding permissible templates must be examined and the best score selected:

$$C(t,1,j) = \underset{1 \leqslant r \leqslant R}{\text{Minimum}} \; C(r,N(r),j-1) + \text{dist}(t,1,j) \quad .\;.\;.\; (2)$$

where R is the number of templates.

Computation proceeds for all templates in parallel in one forward pass through the input pattern. At the end of the input, the score for the best interpretation can be found by examining the scores at the ends of all the templates that are permitted to end an utterance.

Fig. 6 shows some dynamic programming scores computed from the spectrum distance data for the simple example displayed in Fig. 2. The best score for each input frame is displayed as the largest black dot, and the other scores for the same input frame are displayed relative to the best by using smaller dots.

## 3.2 Recording and Using the Word Sequence Information

The above algorithm finds the score for the best time alignment of the best sequence of templates to explain an unknown connected word utterance. However, we are far more interested in the actual sequence of templates which produced this best score, so during the main pass over the input we must also keep track of the word decisions along all the current time registration paths, and then trace them back at the end of the phrase. The information about these word decisions forms a tree structure which "grows new branches" as the unknown input is processed. Fig. 7 shows a simple example of such a "word decision tree" which will be used below. The path between points D and A, for example, corresponds to the sequence of templates T4, T8. There are many paths in the tree, some of which have ended because better-scoring paths have been chosen in the computation of (1) and (2). The tree is currently being extended only at points A, B and C. When the end of the input is reached, the best path and the corresponding best word sequence can be traced back through the word decision tree. Methods for creating this tree structure and recovering the best word sequence are described below.

## 3.3 No Syntax

Vintsyuk [2] suggested the following method of recording template sequence decisions for the simple case in which any of the R templates can follow any other. The method needs three arrays which we shall call T, F and L. The best-scoring template ending at the j'th input frame is defined as

$$T(j) = \underset{1 \leqslant r \leqslant R}{\mathrm{ArgMin}} \; C(r,N(r),j) \quad . \; . \; . \; . \; (3)$$

where ArgMin means the value of the index which minimises the expression. The second array, F(j), records the input frame corresponding to the last frame of the template which precedes T(j). The data structure which corresponds to the word decision tree in Fig. 7 is illustrated for the Vintsyuk method in Fig. 8. (The j'th input frame has just been processed.)

The values in the arrays F and T will be sufficient to recover the best word sequence. After processing the last frame of an input phrase of length M frames, the final template in the best sequence is T(M), the last but one template is T(F(M)), preceded by T(F(F(M))), and so on until the template corresponding to the beginning of the input is reached.

In order to fill the array F, we need the third array, L, which is used to hold "word links". L(t,i,j) holds the number of the input frame corresponding to the end of the template previous to the t'th template, determined along the best time registration path up to the point (t,i,j). In the example in Fig. 9, L(2,i,j) holds the value J2, which is the input frame at which the previous template ended, along the best path to (2,i,j). F(J2) holds the value J1, and T(J2) holds the template number T3.

The word link information in L propagates with the scores, so that

$$L(t,1,j) = j-1 \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; (4)$$

$$L(t,i,j) = L(\, t,i-a,j-1 \,) \quad . \; . \; . \; . \; . \; . \; . \; (5)$$

where a is the index chosen in (1).

For each input frame, the corresponding entry in the array F can now be made, using the value of L from the end of the best scoring template:

$$F(j) = L(\, T(j),N(T(j)),j \,) \quad . \; . \; . \; . \; (6)$$

As with the scores, it is only necessary to store the values of L for the current input frame.

## 3.4 Using Syntax

In many applications of speech recognition there is some knowledge of the permissible order of speaking the vocabulary words. For instance, in data entry to computers the words must normally be spoken in a certain order (e.g. a control word followed by some data) so that the computer can process the information.

The network in Fig. 10 shows an example of a set of word sequence rules (i.e. a syntax) that might be applicable to a subtask in an aircraft cockpit. All utterances accepted by the syntax correspond to routes through the network, e.g. "Waypoint 4", "Height 17434" and "Radio frequency 37.25". Th "silence" template is used to explain the input pattern while there is no speech, and the "reject" template deals with utterances and noises which do not fit the rules. (These techniques will be explained in Sections 4.1 and 4.2.)

We could specify an equivalent syntax (called AIRSYN) in the following form:

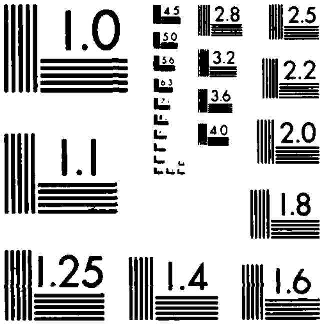
```
DIGIT  = { 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 0 }
WPC    = { waypoint / channel } DIGIT
RF     = radio frequency < DIGIT > { decimal < DIGIT > / }
HT     = height < DIGIT >

AIRSYN = < < SILENCE > { REJECT / WPC / RF / HT } >
```

By including such word sequence information in the recognition algorithm, there are several advantages. It prevents the input from being recognised as a phrase which is just nonsense and, by computing scores only for those sequences of words which "make sense", the amount of computation can be significantly reduced. Of course, if nothing is known about the possible order of the words, the syntax must simply allow that any word can follow any other word.

AD-A132 092 SPEECH PROCESSING(U) ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT NEUILLY-SUR-SEINE (FRANCE) MAY 83 AGARD-LS-129

2/2

UNCLASSIFIED F/G 17/2 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

The method of recording and recovering word decisions can easily be extended to the case of finite state syntaxes (i.e. those that can be drawn as a directed graph, with templates in the arcs). Junctions between templates are called "nodes". Vintsyuk's algorithm deals with a "syntax" containing only one node. When there is more than one node we need to record the template selection decisions made at each node. More information about the JSRU method can be found in Bridle et al. [3].

## 3.5 Continuous Operation

The connected word recognition algorithm described above is suitable for recognising discrete utterances when augmented with an algorithm for determining the beginning and end of a spoken phrase. However, it is not difficult to extend the algorithm to operate continuously. This can avoid the need for explicit endpoint detection (see Section 4.1), and it allows the output of word decisions while the talker is still speaking.

The contents of array L for the current input frame defines those places where the word decision tree is being extended. In Fig. 11, this would be at points A, B and C. The paths back from these points converge at point D. Thus no further input can change the decision that the first two words are T5 and T0, and this decision can therefore be output, even before the end of the utterance.

## 4.0 PRACTICAL CONSIDERATIONS

This section records a variety of techniques and considerations that arose while designing computer programs and specifying a real-time hardware implementation.

## 4.1 Detection of Utterance Endpoints

One of the most difficult problems in isolated word recognition is determining where the word starts and finishes ("endpoint detection"), because this process usually precedes pattern matching. It has been suggested that most errors made by speech recognition machines are basically endpoint errors. Our technique avoids deciding the positions of the boundaries between words before deciding the identity of the words, and the same method can also be used to deal with the utterance endpoint problem. In "isolated utterance" mode we would use syntaxes which start and end with "silence templates", which are simply examples of the background noise spectrum. In principle, it is then only necessary to find the start point very roughly, to move back far enough to include some silence, and to start recognition using the initial silence template. The final silence template can similarly be used to deal with the end of the utterance.

In practice, our real-time machine is normally run in continuous recognition mode, using a syntax which includes one or more silence templates in loops for explaining the input pattern between the spoken utterances. In Fig. 10, for instance, the "silence" template will explain the pauses between phrases, but extra silence templates would be used if we expected the user to pause between words in the phrase.

## 4.2 Non-Vocabulary Words

It is possible to make provision for the speaker uttering words that are not in the vocabulary. In isolated word recognisers, there is usually provision for rejecting an utterance for which the scores of all the templates are worse than some preset threshold. For connected word recognition the corresponding operation is to reject a portion of an utterance while accepting the rest. Our method is to have very elastic "pseudo-templates" which always produce the same "distance" when matched with any input frame. By incorporating these "wildcard" templates into the syntax, spurious inputs, such as breath noises and unknown or out-of-context words, can be matched (and rejected) at selected points in the syntax. The syntax in Fig. 10, for instance, can reject complete spurious utterances, but extra wildcards would be needed to deal with the spurious sounds which could occur between the words within a phrase. The distances for wildcard templates have to be chosen carefully to make it unlikely that a wildcard will be selected when a word in the permitted vocabulary is spoken. If a wildcard is chosen in preference to the correct vocabulary word, it implies that the word has been spoken significantly differently from the version stored in the template.

## 4.3 Acoustic Analysis

The acoustic analysis used in our earlier speech recognition work was based on the JSRU channel vocoder [4], which was designed for low-bit-rate communications. The analysis could therefore be assumed to give a compact description of the speech signal while preserving at least enough information to allow speech communication. In fact the analysis channel filter bandwidths and spacings, and the logarithmic amplitude scale, are all broadly consistent with the usual psychophysical models of the auditory system.

The standard channel vocoder analysis in our computer program produces speech spectra at the rate of 50 frames per second, but the analysis that is currently used in the real-time implementation produces 200 frames per second. This quantity of data, if used directly, would lead to a very large amount of computation (the computation rate is proportional to the square of the frame rate), and to a large template store. A variable frame rate procedure [5] has therefore been adopted, which uses all input frames when the spectrum is changing most rapidly, and omits a high proportion of the frames when the spectrum is relatively constant. Thus the variable frame rate procedure also has the benefit of emphasising the non-stationary portions of the speech sound pattern, where there is probably most linguistic information.

### 4.4 Spectrum Distance Measure

The acoustic analysis data is only used as the input to the calculation of a distance or measure of dissimilarity between two frames. Ideally this distance should not be greatly affected by the loudness of the speech, the background noise or the transmission conditions (e.g. telephone line, position of microphone, etc.), but it should be sensitive to important differences between the shapes of the two spectra. In practice we have used the square of the simple Euclidean distance, but we can also incorporate additional processing to make some adjustment for both amplitude variations and background noise.

The vocoder analysis has a limited dynamic range, and in speech with a good signal-to-noise ratio the spectrum in the gaps between words is reasonably constant. As described in Section 4.1, we use a "silence template" to account for these gaps between words, and this normally matches well. In realistic conditions, the background noise can have a relatively high level and an arbitrary spectrum shape and can also be varying. If the background noise is varying, we need to estimate the background noise continuously and to modify the silence templates accordingly. However, background noise also affects the fit of all the word templates, and therefore we include some additional processing which reduces the effect of the background noise on the distance measure.

Klatt [6] proposes a number of improvements to a simple filter bank analysis. In Klatt's noise compensation method the distance computation needs four spectra: input speech, input noise, template speech and template noise. In our method we first compute a weighting function for each speech spectrum, based only on that spectrum and the current estimate of the noise spectrum. The spectrum distance calculation combines the two speech spectra with their weighting functions in a way that makes full use of the original information and provides a measure of the amount of difference between the underlying speech spectra. The method leads to a particularly efficient hardware implementation because the weighting function can be represented using very few bits, and because the distance calculation can be pipelined.

### 4.5 Computation and Storage

The notation used above implied that the scores (C) and the word links (L) need three-dimensional arrays to store them, but because of the order in which the processing is done there is no need for the input frame index (j). Consequently, the implementations store one score and one word link for each template frame. Similarly, spectrum distances are always between a given template frame and the current input frame.

Compared with an isolated word recogniser with the same size vocabulary, using the same methods of pattern representation and dynamic programming matching algorithm, the above connected word recognition algorithm takes about the same amount of computation per input frame, although it needs a greater amount of working storage. Because all words must be considered in parallel, the working storage for the scores is increased by a factor of the vocabulary size. The word links L and the arrays F and T also need extra storage, but the total amount of working storage is less than that required to hold the template patterns themselves.

The most computationally-intensive operation is the calculation of the spectrum distances, but this is a very regular operation which is well suited to special, high-speed circuitry. The spectrum distance calculation needs rather rapid access to the template data.

### 4.6 Score Pruning and Scaling

The amount of computation can be reduced significantly by pruning the dynamic programming scores (Lowerre's "Beam Search" [7]). For each input frame, all scores which are more than some specified distance away from the best score for that input frame are removed from further consideration. This avoids considering relatively unlikely interpretations of that part of the input pattern which has already been processed, but keeps the options open if there seems to be ambiguity. Our present computer program and the real-time equipment also reduce the range of numbers needed to represent the scores by setting the best score for each input frame to zero, by suotraction. Fig. 6 can be regarded as a representation of the actual, modified scores, with the white areas corresponding to "pruned" scores.

## 5.0 LOGOS - A SPEECH RECOGNITION MACHINE

Although our connected word recognition algorithm has been available for some years in a non-real-time computer program, our experiments have been limited by computation speed and insufficient memory to store more than a dozen templates. To enable the pattern matching technique to be explored further, a powerful and flexible real-time equipment, based on our algorithm, has been designed and constructed under contract by Logica Ltd. This equipment, known as "Logos", offers applications research laboratories the capability to evaluate the use of this type of speech recognition in many applications.

The main aim was to achieve flexibility and performance, at the expense of storage and computation costs. An important requirement was to allow interaction between the recognition process and the software which makes use of the recognition results. In one direction, the constraints of the application can be used to guide the recognition process, by using syntax and setting various parameters. In the other direction, the output of the recogniser can include much more than just the identity of the templates in the best-fitting template sequence. Other information which has been made available includes durations and scores for the intervals of the input which are "explained" by each template, and some indication of alternative (sub-optimal) template sequences, which might be chosen by the application software if they "make more sense".

A functional overview of Logos is shown in Fig. 12. The acoustic analysis section includes a 19-channel filter bank analyser with much better amplitude and time resolution than the vocoder, plus a microprocessor (the Front End Processor). The front end processor implements various transformations of the raw filter bank spectrum cross-sections, including variable frame rate, background noise spectrum estimation and the first stages of the noise compensation algorithm. There is a buffer of several seconds duration between the acoustic analysis section and the pattern matching section, to allow for variable computation rates. The main computational load is in the distance calculation, which is handled by a high-speed special-purpose hardware module. The spectrum distance includes the noise compensation technique referred to in Section 4 above. Up to 16 dedicated microprocessors share the work of the dynamic programming steps at the heart of the algorithm, while the control processor keeps track of the syntax and word decisions. This organisation exploits the fact that the calculations for each template are substantially independent of the calculations for the others. Interactions between template processing only occur at the beginning and end of each template and in the score pruning. All the microprocessors are Intel 8086's. More information can be found in Peckham et al. [8].

Logos has been designed to recognise sequences of connected words using a vocabulary of up to 200 templates. The number of templates that can be stored in the machine is only limited by the amount of template memory, and the length of the templates. Finite state syntax with loops may be specified to guide the recognition process, which can cope with an average of about 100 words "active" at any one time, depending on the number of dynamic programming processors. The wildcard facility (Section 4.2) allows the equipment to deal with non-vocabulary words, and non-verbal noises such as coughs and breath, at selected points in the syntax. The recognition of key words can control the switching of syntaxes or choice of vocabulary. Continuous operation (Section 3.5) permits Logos to handle utterances of any length.

In common with most currently available recognisers, Logos is essentially speaker dependent, requiring each user to provide example utterances of each of the vocabulary words. Training of the machine typically requires the user to speak each of the vocabulary words at least once in isolation. The extraction of templates is done using the recognition algorithm, with a suitable syntax consisting of wildcard and silence templates. By specifying a more complicated training syntax, it is also possible to extract templates embedded in known carrier phrases [9]. Logos has the ability to store and retrieve templates from a host computer. For research purposes, the attached host may be placed in complete control of Logos, when monitoring of recognition performance and acquisition of intermediate results are possible.

## 6.0 CONCLUSIONS

The two main features of our current approach to connected word recognition are the use of whole word templates and the one-pass dynamic programming algorithm for deciding the identity of the words spoken. It is a simple-minded, brute-force approach and its performance falls far short of that of a human listener, even when the machine is set up to suit the speaker. However, we believe that machines based on these principles will be useful in many voice input applications, particularly for trained operators of complex machines.

These methods can form a stepping stone for future research in automatic speech recognition. There is the potential to improve the whole word pattern-matching approach by incorporating more information about the words in each template. This could include information about permitted timescale distortion [10], and about the variability of the spectrum at each frame. Even for quite different approaches to automatic speech recognition, the one-pass dynamic programming organisation can be an inspiration as a method for analysing and "decoding" the speech.

The experiments which we have carried out so far using whole word pattern matching have been limited, and only small vocabularies have so far been tested. The power and the limitations of these methods have yet to be ascertained and the real-time equipment that is now becoming available will help to explore their potential. Possibly more important is the need to explore the consequences of using automatic speech recognition in many different application areas. It is hoped that the present generation of speech recognition equipments will be used to find suitable ways of using speech recognition in complete systems, so that when smaller, cheaper and perhaps better speech recognition equipments become available in the future, they can be used profitably without further delay.

## 7.0 REFERENCES

1. Bridle, J.S. and Brown, M.D., "Connected word recognition using whole word templates", Proc. Institute of Acoustics, Autumn Conference, Windermere, 1979, pp.25-28.

2. Vintsyuk, T.K., "Element-wise recognition of continuous speech consisting of words of a given vocabulary", Kibernetika (Cybernetics), No.2, 1971.

3. Bridle, J.S., Brown, M.D. and Chamberlain, R.M., "A one-pass algorithm for connected word recognition", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Paris, 1982, pp.899-902.

4. Holmes, J.N., "The JSRU Channel Vocoder", Proc IEE., Vol.127, Part F, No.1, pp.53-60, February 1980.

5. Bridle, J.S. and Brown, M.D., "A data-adaptive frame rate technique and its use in automatic speech recognition", Proc. Institute of Acoustics, Autumn Conference, Bournemouth, 1982, pp.C2.1-C2.6.

6. Klatt, D.H., "A digital filter bank for spectral matching", Proc. IEEE, Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, 1976, pp.573-576.

7. Jelinek, F., "Continuous speech recognition by statistical methods", Proc. IEEE, Vol.64, No.4, pp.532-556, April 1976.

8. Peckham, J.B., Green, J.R.D., Canning, J.V. and Stephens, P., "A real-time hardware continuous speech recognition system", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Paris, 1982, pp.863-866.

9. Rabiner, L.R., Bergh, A. and Wilpon, J.G., "An embedded word training procedure for connected digit recognition", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Paris, 1982, pp.1621-1625.

10. Russell, M.J., Moore, R.K. and Tomlinson, M.J., "Some techniques for incorporating local timescale variability information into a dynamic time-warping algorithm for automatic speech recognition", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Boston, 1983.

## 8.0 ACKNOWLEDGMENTS

This text draws on work with several colleagues at the Joint Speech Research Unit and at Logica Ltd.

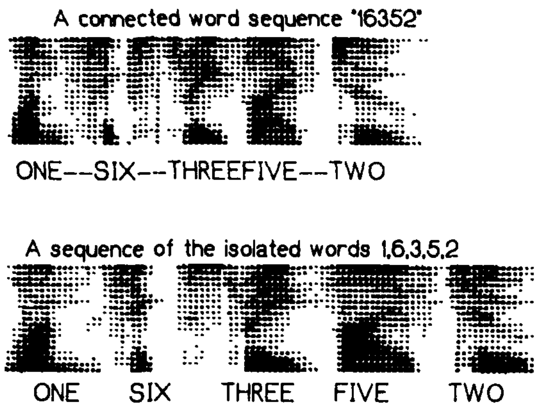


Fig. 1. Sound patterns of an input utterance and a template sequence.

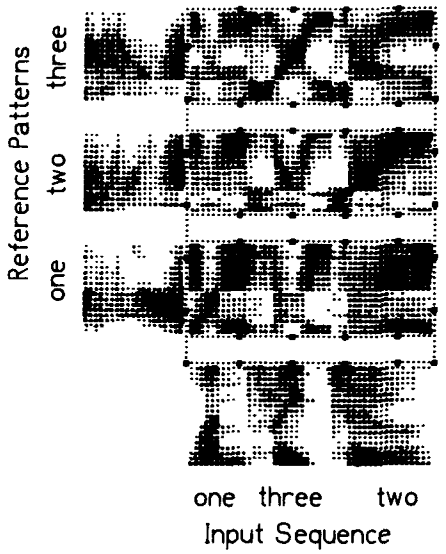


Fig. 2. Spectrum distances for a simple example.

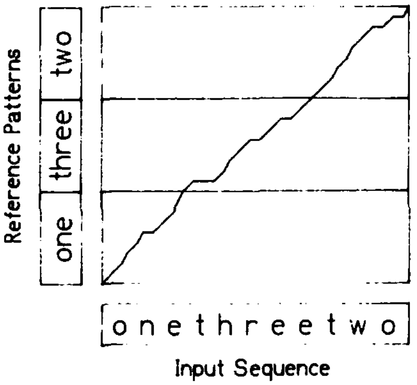


Fig. 3. Time alignment of a sequence of templates to a connected word input.

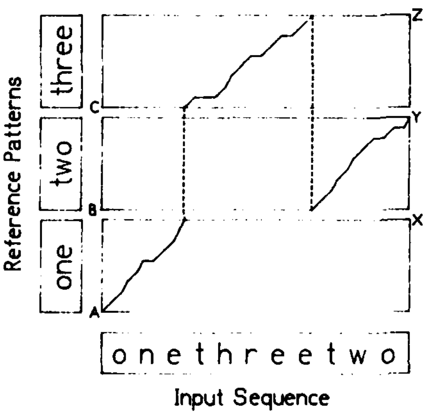


Fig. 4. An alternative representation of the time alignment.

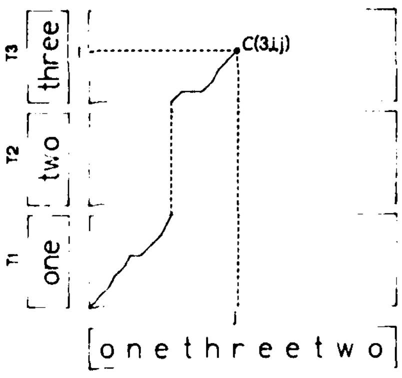


Fig. 5. A connected word recognition score.

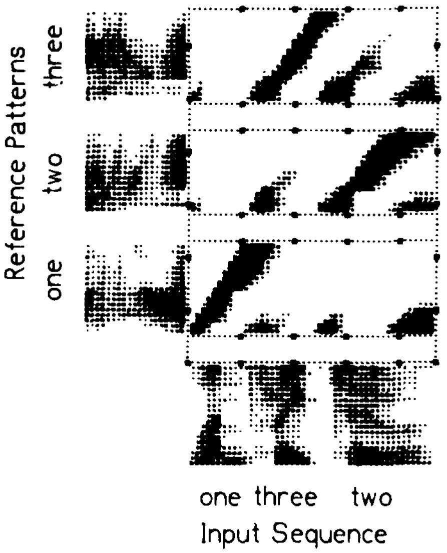


Fig. 6. Dynamic programming scores for a simple example.

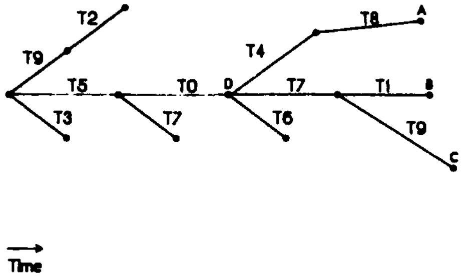


Fig. 7. Word decision tree generated during recognition.

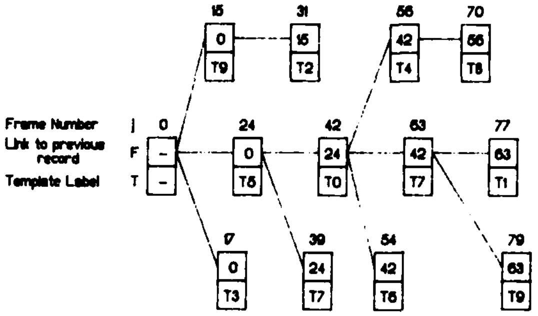


Fig. 8. Structure of Vintsyuk's word decision data (indexed by input frame number).

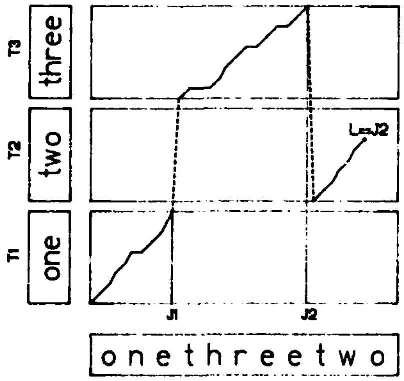


Fig. 9. Propagation of the word link information.

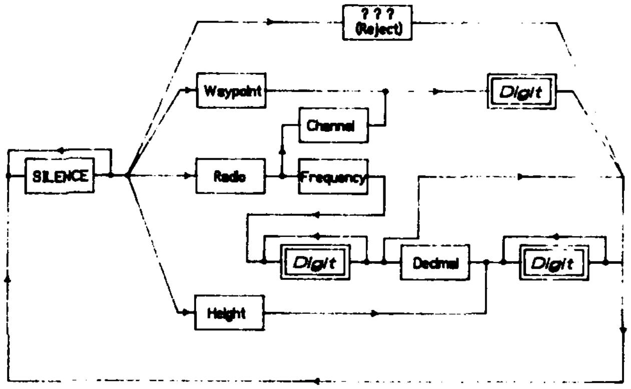


Fig. 10. An example of a finite state syntax with loops.

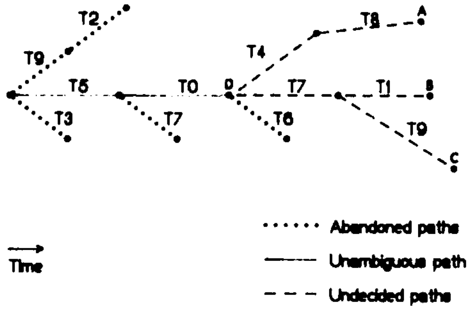


Fig. 11. Properties of the word decision tree which are used in continuous operation.

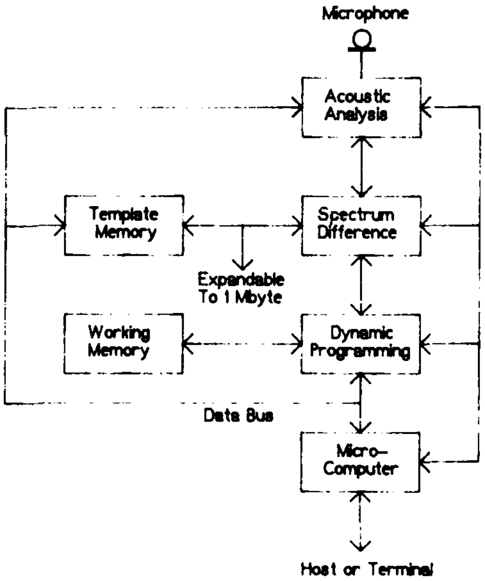


Fig. 12. Functional overview of 'Logos', a real-time continuous connected word recognition system.

# VOICE TECHNOLOGY IN NAVY TRAINING SYSTEMS

Robert Breaux*
Naval Training Equipment Center
Orlando, Florida 32813

MaryAnn Blind**
Northrop Services, Inc.,
and Robert R. Lynchard**
Advanced Technology, Inc.
Orlando, Florida 32803

This paper is designed to familiarize those involved in training development with the nature, constraints and applications of computer voice technology (CVT). It will also show you how to evaluate voice technology for meeting training requirements, and how to incorporate CVT into your training design.

Let's look at the technology - the "how-does-it-work" of computer speech generation and voice recognition. This will not be a highly technical discussion for two reasons. First, a technical discussion would require a highly technical background and would not further your use of the technology. Second, the technology is continually diversifying. It is becoming increasingly difficult to keep track of the various agencies and vendors involved in CVT, no less their individual approaches, techniques, and special interests and applications. Rapid advances in language analysis and other related technologies add to the rapid technical growth of this field.

## Computer Speech Generation (CSG)

Initially, we'll look at computer speech generation (CSG) which is a simpler technology to begin with than voice recognition. There are, essentially, three types of CSG - digitized, word generated, and phoneme generated (see Figure 1).

## Digitized Speech

Sounds are produced as wave forms. When a word is entered into the computer, the wave form is digitized and becomes a pattern. This pattern represents all of the stress, pitch, and pause associated with the word. When the complete word pattern is stored, the speech that is generated from it is called digitized, and the resulting sounds are very real and natural. Storing complete word patterns, however, requires a lot of memory.

## Synthesized Speech - Word Generated

In order to save computer memory, the digitized pattern can be compressed, and then may be stretched out again to be generated. When this is done, however, less of the word pattern is stored, the sound wave becomes slightly distorted, and the results sound somewhat metallic or mechanical. This speech is called "word generated" and is one type of synthesized speech - speech which sounds somewhat synthetic rather than natural.

## Synthesized Speech - Phoneme Generated

Another type of synthesized speech is phoneme generated, which uses the least amount of computer memory. Phonemes are the basic, or smallest, phonological elements of speech. All of the words in most languages can be formed from combinations of phonemes. Word patterns are formed in two ways. The digitized phonemes for a word can be selected by the computer either according to pronunciation rules or by preprogramming sequences of phonemes which are combined to form a word pattern. This pattern is stretched out to real time into a word sound wave. The size of the available vocabulary is limited only by programming requirements. Phoneme generated speech is more distorted than word generated speech since the effect sounds have on one another when spoken together is not accounted for. This is the speech that sounds robotic.

* This paper was developed from a portion of the effort completed under Government Contract N61339-80-G-0003 with Eagle Technology, Inc. Opinions expressed herein are those of the authors and do not necessarily reflect official policies of the United States Navy.

** Dr. Blind and CWO2 Lynchard, USN (Ret.) were employed by Eagle Technology, Inc., at the time that this paper was prepared.

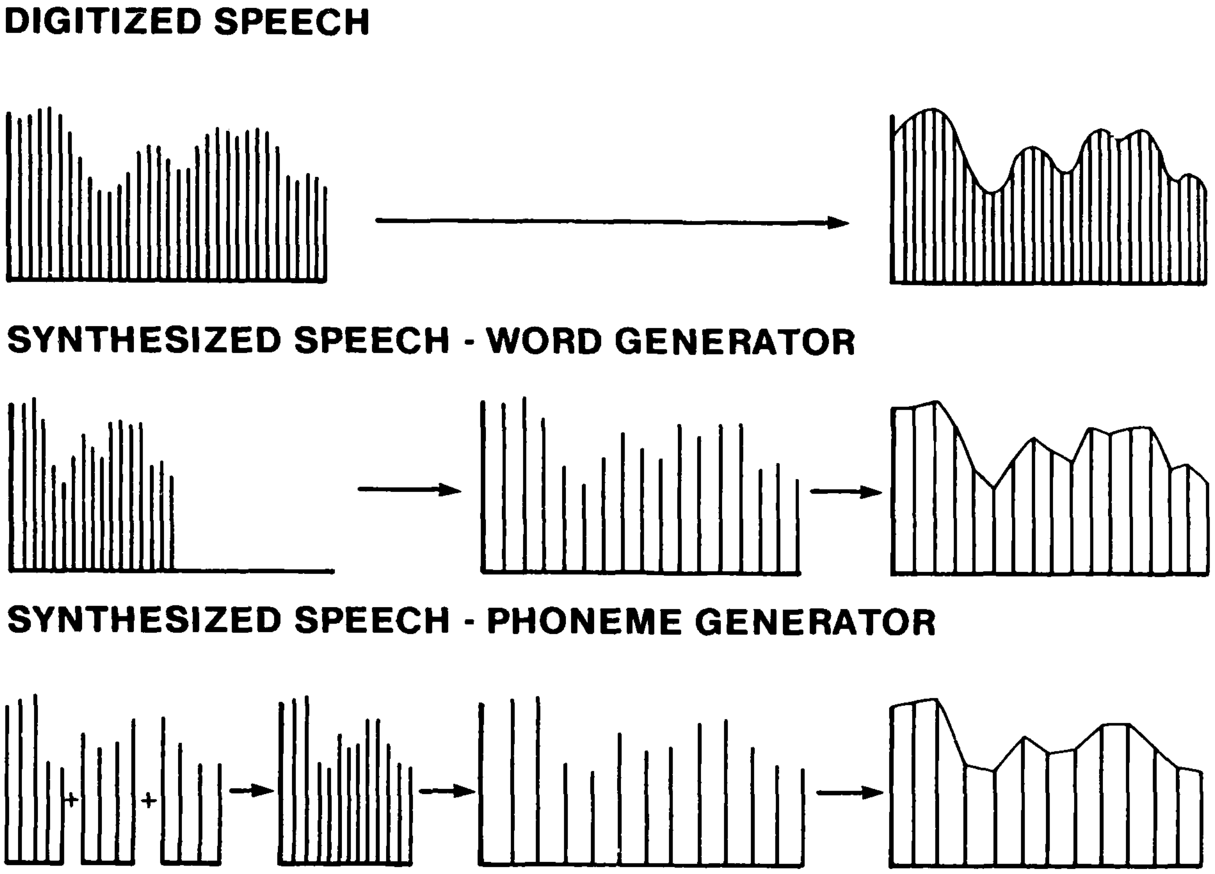

DIGITIZED SPEECH
SYNTHESIZED SPEECH - WORD GENERATOR
SYNTHESIZED SPEECH - PHONEME GENERATOR
+
+

Figure 1. Characterization of Computer Speech Generation.

## CSG Characteristics

By looking at these different sound waves, you can see the quality difference computer memory can make. Phoneme generated speech can be improved through the use of filters and other technical advances which smooth out the sound wave. This speech is the most flexible, offering an unlimited vocabulary at low cost. Phoneme generated speech also can provide multilingual vocabularies.

Word generated speech offers a limited vocabulary, depending on the available computer memory and the amount of compression used. The more compressed the word patterns are, the lower the speech quality. Word generators use up to five times the computer memory of phoneme generators and are equivalent in cost to phoneme generators as they trade off vocabulary size for speech quality.

Digitized speech is much more sophisticated than synthesized and uses at least twice as much memory as the highest quality word generated speech. Digitized speech can use as much as ten thousand bits per second of storage where phoneme-generated speech can use as little as four hundred bits per second of storage. Digitized speech requires a large computer capability resulting in high quality, natural sounding speech and a large vocabulary at much higher costs.

A choice of which speech generation method to employ depends upon your specific application and involves certain trade-offs between vocabulary size, speech clarity and cost in terms of computing power.

## COMPUTER VOICE RECOGNITION (CVR)

Now let's discuss computer voice recognition (CVR). A computer can be controlled by verbal commands through the addition of a voice recognition capability. This type of system consists, basically, of a recognition capability, a computer, necessary interfaces, appropriate software, and the programs to incorporate the recognizer. Different manufacturers produce these systems using different technologies. Some build computers with the voice component built in, and some produce voice units which can be interfaced with certain computers. Currently, research is ongoing to produce a voice recognizer on an electric component or "chip."

## CVR Process

Generally, computer voice recognition works in this manner. As in speech generation, sounds are digitized and made into patterns by the computer. The computer has a set of patterns stored in its memory, which it compares to the incoming voice patterns. If the computer can find a close enough match according to preset probability parameters, it will select the correct item. If not, it will not recognize the pattern. In most systems the parameters can be changed by the user. How to change them is a part of the design issue which is covered in more detail later.

The comparison process is referred to as pattern matching.[1] Most voice recognizers currently employ pattern matching on the whole word, which requires storage of whole word patterns and the matching of the pattern of each incoming whole word within some probability parameter. Some research topics in pattern-matching are as follows:

- Word spotting, where attention is focused on the portions of speech that distinguish words.
- Matching short-time spectra, which matches the sequence of small units with similar stored units.
- Matching spoken phonetic sequences with stored phonetic sequences for all possible words.

On some occasions the computer may misrecognize a pattern and respond accordingly. Words which sound alike can cause this problem, for example "run" and "one." If one of these words cannot be eliminated from the program, then the computer can be taught a different pronunciation for one of the words. In this case "run" can be entered as "execute." When establishing a word set, the vocabulary should be checked for words that can easily be confused. These words should be replaced, if possible.

On other occasions the computer may not recognize a word at all, due, usually, to voice changes which result from stress, fatigue or illness. In this situation additional voice patterns need to be entered into the computer.

## Training The Computer

Entering voice patterns into the computer is referred to as "training the computer," or utilizing voice data collection and enrollment, and consists of inputting repeated entries of the words which are to be recognized. The computer averages the entered patterns into a representative pattern for each word. When non- or misrecogni- tion occurs, the computer must be "retrained," and the additional inputs are averaged into the existing pattern. The type of training required by a system depends on whether the system is speaker dependent or speaker independent.

## Speaker Dependence/Independence

Speaker dependent systems are specifically tailored to an individual's speech patterns. Each speaker must train the system to recognize the speaker's voice by repeating each word in the vocabulary from one to ten times, depending on the requirements of the particular system. This process results in a system that accepts variations in pronunciation. If the computer has difficulty with non- or misrecognition of any word, the speaker can usually correct this with a brief retraining of the computer on that word. Non- or misrecognition can result from voice changes due to a cold, the time of day, or other circumstances.

Speaker dependent systems are primarily useful when the speaker will use the system to repetitively perform a task.[2] This type of system is also useful when the task requires that only a properly qualified person perform that task, such as in quality assurance, production cost accounting, or electronic funds transfer. Many systems, however, have the capability of storing several speakers' sets of speech patterns, so that any individual speaker can call up his or her patterns into the computer and use the system with little or no retraining. Different speakers' patterns can also be stored on disc or tape and loaded as needed. This is useful when a system is being used to train persons in a task, or when several persons perform a job at different times, such as in shift work.

Speaker independent or universal systems are designed to recognize the voices of persons for whom no previous voice samples have been supplied. In experimental systems, this function is accomplished by constructing a dictionary of reference patterns that model the peculiar speech characteristics of each word. Commercial systems incorporate independent speaker recognition by constructing a very large data base from hundreds of speakers in order to appropriately model all of those with diverse patterns. In sophisticated systems of this type, the patterns may be adapted automatically for better recognition performance as a new speaker continues to use the system.[3] These systems generally are less accurate, operate with smaller vocabularies than speaker dependent systems and, in practice, may not work for all speakers. While universal recognition is a desirable goal for speech recognition, the loss of accuracy, the additional complexity of the recognition logic, and the fact that not all people will be understood generally limit its use to only those systems which are designed for public use. In those situations it is not reasonable to expect users to train the system to recognize their individual voices.[4] Depending on the use of the system, however, it is possible for a speaker who is having recognition difficulty to enter a word pattern or two, which are then averaged into the data base for those words.

Speaker independent systems are required when the operator is a casual user, and there is no requirement to validate that the operator is authorized to perform a task, such as finding a telephone number, or the flight weather between two locations.[5] Speaker independent systems can also be used in conjunction with live operators for such things as phone banking where the live operator may first need to verify the user.

## Isolated Word/Connected Word Recognition

Most systems available today recognize isolated word speech, which requires a distinct pause by the speaker between each word or utterance and no substantial pauses within words. An utterance is defined as a word or sequence of words restricted in time to between one and one-half and three and one-half seconds, depending on the system. Isolated word recognition systems are useful when a large number of people must have access to the system, and when a single response can be given to a computer question or prompt. For other purposes, however, this speech is awkward and unnatural; connected speech may be a more viable alternative.

Connected word recognition systems can recognize a sequence of words spoken in natural cadence and provide for faster data entry. Connected speech has such applications as warehouse routing and inventory control, and zip code entry for mail sorting systems. These applications are restricted, however, and recognition of unrestricted or continuous speech is still experimental. Continuous speech recognition is discussed later within the topic of speech understanding.

The speed of speech is more of a factor in connected speech than isolated word recognition. Usually, the length of the sound pattern is averaged over repeated entries. However, another method, called dynamic programming, adjusts the time bases of stored sound patterns to those of an unknown utterance, resulting in higher recognition accuracy.

At this time connected speech recognition requires ten times as much computation as isolated word recognition, and consequently, has higher costs. Further research can bring these costs down.

## Vocabulary Size

A third primary characteristic of voice recognizers is vocabulary size. What constitutes small or large vocabularies will vary depending upon the context and the state-of-the-art. For example, an isolated word, speaker dependent recognizer with an utterance capacity of three hundred words or more may be considered large. However, the vocabulary of a connected speech, speaker independant recognizer may be considered large if it is greater than fifty words.[6] Connected word vocabularies can contain as many as one hundred and twenty words.

The total vocabulary size for isolated word recognizers usually ranges from twelve to one thousand or more words, although at each word position there are typically ten or fewer words in the active vocabulary set from which the correct word must be chosen. For each task a syntax tree effectively determines which subsets of the entire vocabulary are active or available at different stages within the task procedure sequence according to a "menu" format (see Figure 2). As the number of possible word choices increases, the computer's ability to discriminate between the correct word and the incorrect words decreases.[7]

Vocabularies for connected speech recognizers are limited due to the added complexity of sorting words out of sequences. This requires more complex acoustic discriminators. Some connected speech recognizers use sentence patterns where the sentence types are pre-programmed as in the IF-THEN statement, reducing the number of options for any position in the sentence (see Figure 3).

Vocabularies may be expanded to an almost unlimited size through the use of additional data storage and branching techniques. For example, (see Figure 4) in a recognizer with only a five word recognition capacity, any one or all items may be designated as "control words." The use of a control word would cause the system to branch to another program that contains a different set of five words any of which may again be control words.[8] The use of any of these techniques depends specifically on the training requirement.

## Speech Understanding

One step beyond voice recognition is speech understanding, which requires that only enough of the key words in an utterance be correctly decoded for an appropriate response to be elicited from the computer, such as the retrieval of some stored

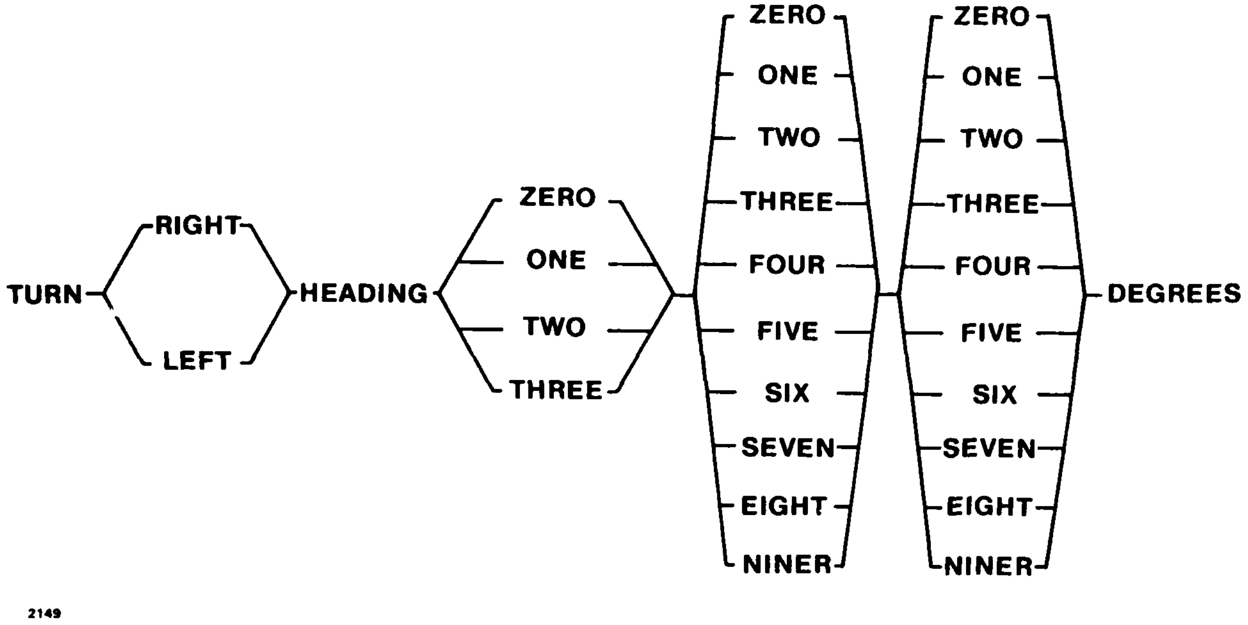


Figure 2. Syntax Tree for Phrase "Turn Left Heading One Zero Six Degrees."

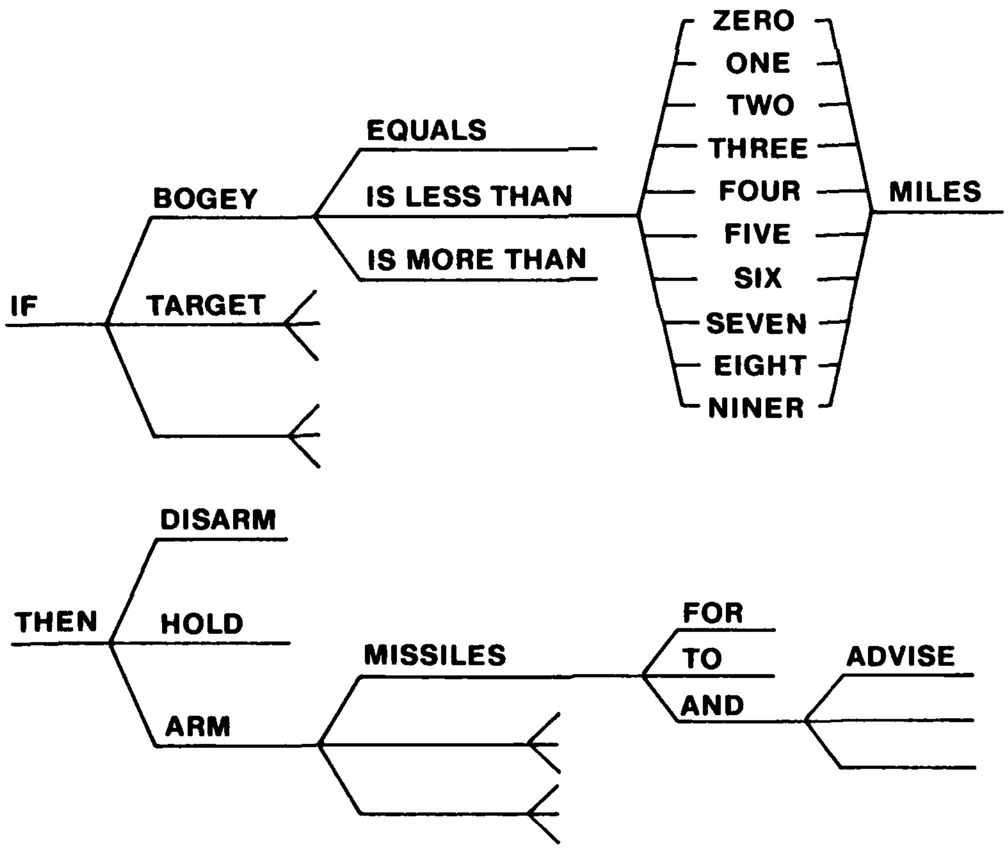


Figure 3. Pattern Format for Sentence "If Bogey is Less than Four Miles, Then Arm Missiles and Advise."

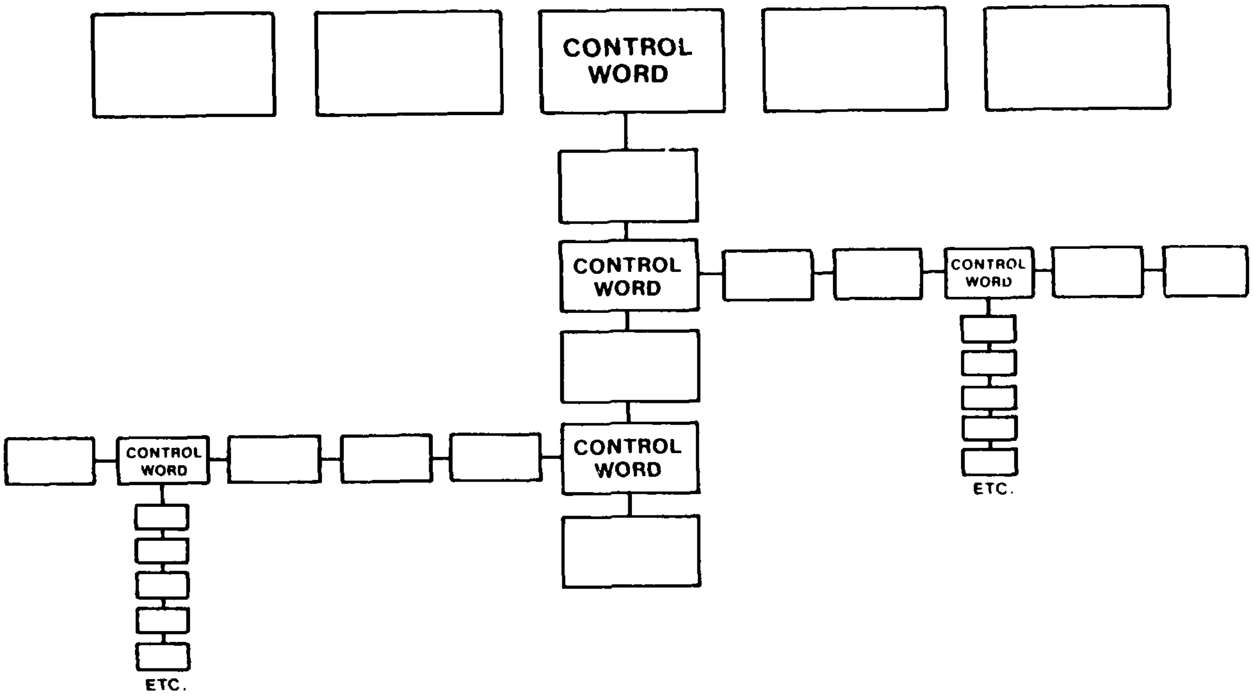


Figure 4. Branches on Control Words to Five Word Vocabularies.

information.[9] A further step in speech understanding would enable the system to employ knowledge of the real world and of human speech to understand speaker intent. This type of system requires significant advances in computer technology, natural language processing, and artificial intelligence. Artificial intelligence has been defined as "that part of computer science that is concerned with the symbol manipulation processes that produce intelligent action."[10]

## TRAINING APPLICATIONS

Four military applications for training systems will be addressed here which incorporate computer voice technology. In the first three systems, the student would be trained in some form of air traffic control. The student must learn to process visual and auditory information rapidly while making verbal advisories to the aircraft involved. Two of these systems, PARTS and LSOTS, provide control instructions for aircraft during landing operations. The AIC system provides control during tactical maneuvers. With the fourth system, AIDS, the student would be trained in aircraft operations where both the aircraft and the trainer would incorporate computer voice technology.[11]

## PARTS

The Precision Approach Radar Training System (PARTS) for air traffic controllers represents the culmination of work begun in 1972 for the Naval Training Equipment Center. This system was originally called the Ground Controlled Approach Controller Training System (GCA-CTS).

The earliest work on this system identified the PAR control task as an ideal test bed for research in computer voice technology. PAR control is primarily a verbal task not previously amenable to automated training. The vocabulary used is rigidly defined and highly stylized and is, therefore, potentially recognizable by the isolated phrase recognition technology. Performance of the PAR control task requires interaction with a pilot, pattern controller, and tower controller (see Figure 5). This situation is also ideally suited to the development of models for these positions incorporating speech generation.

A series of laboratory studies involving the development of a preliminary training system led to the development of an experimental prototype system. This system was used to demonstrate the feasibility of:

- Employing the automated speech technologies in an operational training environment.

- Developing a training methodology incorporating instructorless training and automated speech technologies without compromising training effectiveness.

- Developing an instructor model which could provide automated adaptive training for a primarily verbal task.

- Devising a performance measurement scheme which would enable the system to provide instructive feedback to the trainee, progress information to the learning supervisor, and input to the instructor model which would enable automated adaptive problem selection.

- Devising techniques for providing the feedback to the trainee and learning supervisor.

- Developing useful models of the verbal and motor behavior of the other persons with whom the precision approach radar (PAR) controller interacts, namely the pilot, pattern controller, and tower controller, as well as a model of PAR controller behavior.

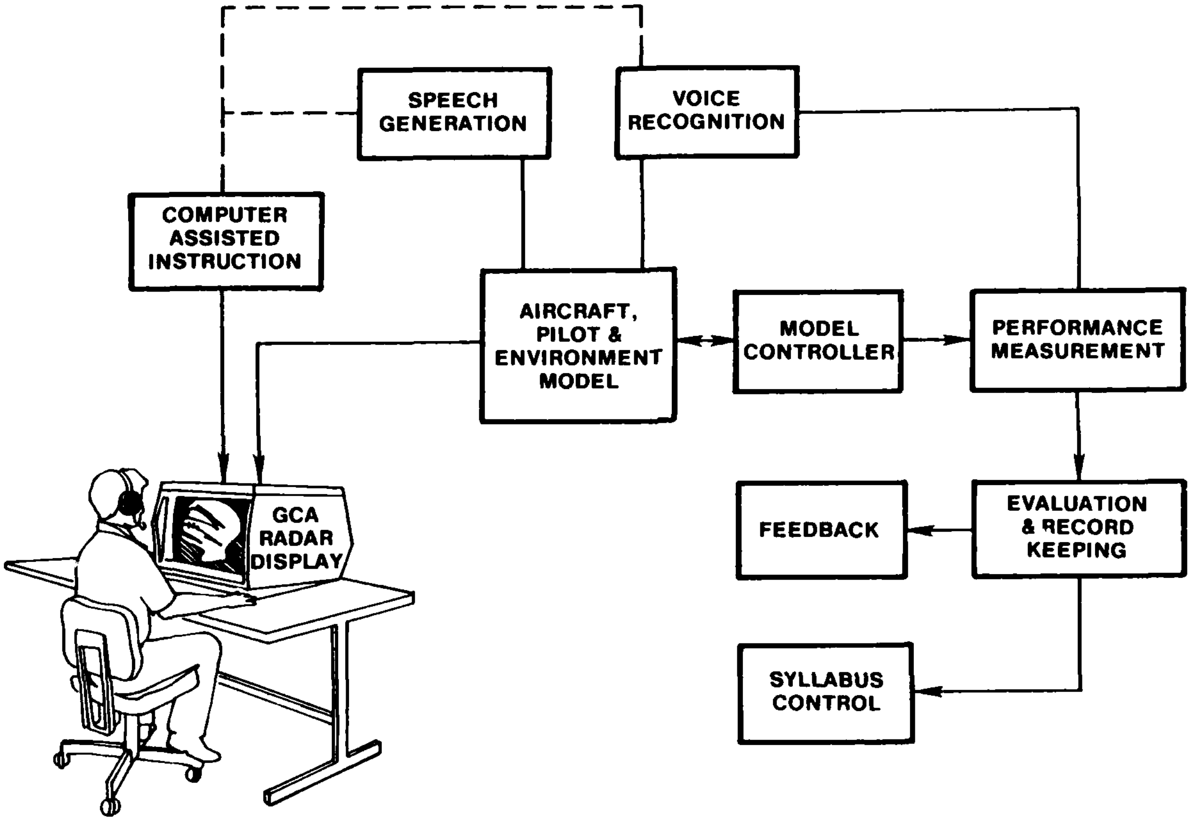


Figure 5. PARTS System Components.

The primary performance requirements under which PARTS was developed include:

- The state of the art in speaker dependent, isolated phrase voice recognition technology.

- Good voice recognition in real time over a relatively large vocabulary containing many similar phrases.

- System visibility by persons not previously trained in the use of voice recognition equipment.

- Training in the PAR control task equivalent to that provided in the existing training environment and in an environment with minimum instructor intervention.

- Realistic stimuli such as radar displays, servo controls, and communications equipment to facilitate transfer of training.

- Training of the student to proficiency within present time constraints.

The resulting PARTS is a stand-alone, experimental, prototype training system. PARTS provides automated, individualized instruction in techniques for providing ground-controlled approaches. In addition, it ᵖrovides a realistic environment in which radar control skills can be practiced un ᵋ. the supervision of an automated instructor. It also provides objective performance measurement and feedback in the form of performance summaries and annotated replays. Although the order of topic presentation is rigidly defined in the basic syllabus, problem difficulty is adapted, amount of practice is varied, and remedial exercises are selected to automatically adapt the basic course to the needs of the individual trainee. One of the major benefits of the system is that it relieves the trainee of the need to devote part of his or her time to serving as a pseudo pilot for other trainees, a requirement when using the existing training device. It also provides enrichment topics for those students who complete the basic course quickly. This provides students who quickly attain the minumum requirements to qualify as air traffic controllers with the opportunity to continue with advanced training topics. Finally, the system provides the learning supervisor with informative feedback regarding the individual trainee's performance.[12] An evaluation of PARTS is reported by McCauley.[13]

AIC System

Research to develop a voice recognition and speech understanding system was initiated in 1977 to support an automated training system for Air Intercept Controllers (AICs). The task of the AIC is to direct an intercept aircraft in a combat situation to destroy an enemy aircraft (see Figure 6). The AIC must make split-second decisions and have rapid, accurate motor responses in controlling the aircraft displayed on a complex monitor. The research involved developing a voice recognition subsystem utilizing new recognition techniques. This system was subsequently tested in a laboratory AIC training model which was under concurrent development.

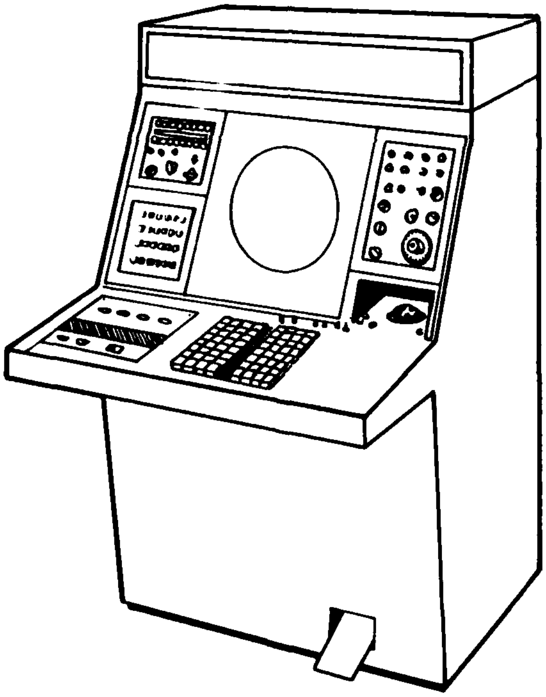


Figure 6. AIC Work Station.

The objectives of the research were to:

- Achieve a greater understanding of the demands placed on computer-based voice recognition by an automated AIC training system.

- Determine if a previously demonstrated limited connected voice recognition system could be effectively combined with the more usual isolated word recognition systems to satisfy the AIC requirements.

- Provide an applications environment for the continued development of speech recognition algorithms to provide focus and ensure the earliest possible realization of an operationally useful recognition capability.

The primary constraints under which the AIC training system was developed include:

- The innovative use of new technologies. The automated AIC training problem represents a significant advance beyond the PARTS in both the application of the voice technologies as well as training systems design.
- A requirement for recognizing a large amount of numeric strings. Isolated phrase recognition will not meet the recognition requirement of AIC training, and a complete connected speech capability is beyond the current technology. Therefore, a mixed isolated phrase and limited connected recognizer was required.

The prototype was a stand-alone system which was used for fleet evaluation for further revision and determination of logistic requirements.[14] An evaluation of the AIC is reported by McCauley.[15]

LSOTS

A Part Task Landing Signal Officer (LSO) Waving Concept Trainer was an exploration in training the conceptual as opposed to the perceptual portion of the LSO's task of controlling aircraft landings aboard a carrier (see Figure 7). The LSO must visually evaluate an aircraft's position on its approach to landing. The LSO assesses aircraft approach and recovery conditions, directs pilot corrections, and advises superiors of recovery feasibility, efficiency, and safety. The LSO is responsible for guiding the aircraft to a safe landing or waving it off for another attempt. The LSO uses a handset to communicate verbally with the pilot and a hand device for activating "wave off" and "cut" lights, requiring the use of both hands. This is a critical task requiring fine discriminations, quick decisions, and eye-mouth coordination. The part-task system teaches the latter two functions.

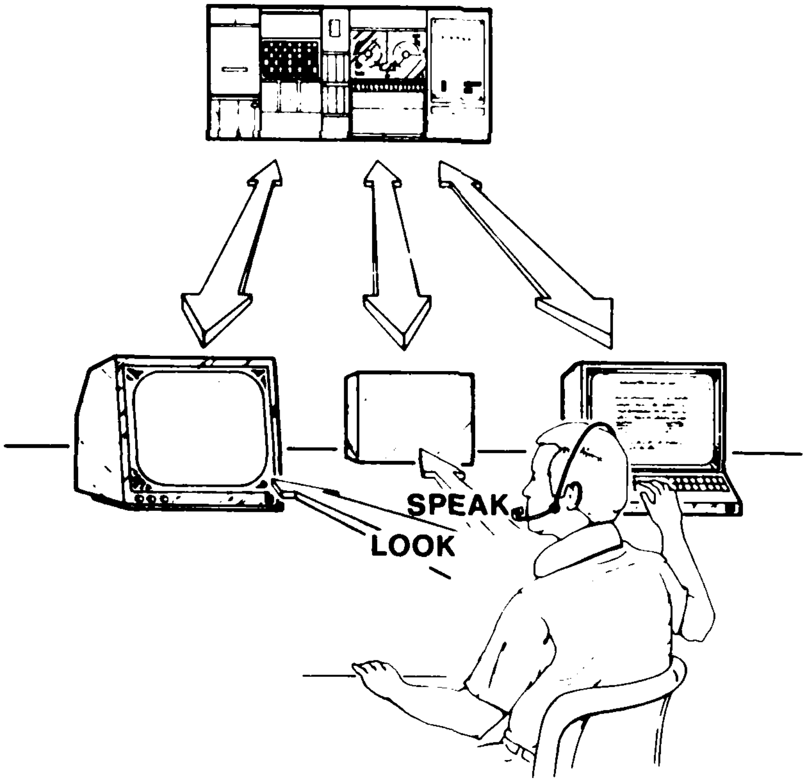


Figure 7. Characterization of the LSO Part Task Waving Concept Trainer.

LSO training incorporates a large percentage of on-the-job training (OJT), and is, consequently, dependant on OJT oppportunities. Reduced operational training opportunities have created a severe manpower shortage of LSOs. A review of the required skills determined that the training of LSOs could be accomplished using automated training.[16] The proposed LSO Training System (LSOTS), based on the part-task system plus a high fidelity visual system, visually simulates the view of the aircraft from the LSO position on the aircraft carrier (see Figure 8). The aircraft moves in response to the student's verbal commands. The system uses computer generated imagery to present the scenario, a voice input, and a keyboard CRT to present instruction and performance evaluation information.

AIDS

The Advanced Integrated Display System (AIDS) is a test bed cockpit trainer for the F-18 aircraft, presenting an example of the use of voice recognition and speech generation in an operational environment which requires the same use in the trainer. AIDS uses voice recognition for control of radio frequencies, TACAN and SIF, and speech generation for verification of commands. For example, the pilot can vocally change the radio frequency, and the computer will vocally verify the new frequency. This system holds potential for gaining information from the computer. The pilct could ask the computer to alert him when the altitude is below five hundred feet. The computer would verify the input by repeating it. Then when the aircraft descends below five hundred feet, the voice generator would notify the pilot. A more complex example has the pilot telling the computer to arm the missiles when the bogey is within ten miles, and to report accomplishment to the pilot. The computer would verify the input, track the bogey to within ten miles, arm the missiles, and notify the pilot.

Computer voice technology is perceived as having a high potential for military training applications. This is particularly true in the training of interactive skills for teams where the computer can model team members and provide performance measurement. Improved procedures for team task analyses have made it possible to model the team, thus providing a base for adaptive training techniques.[17]

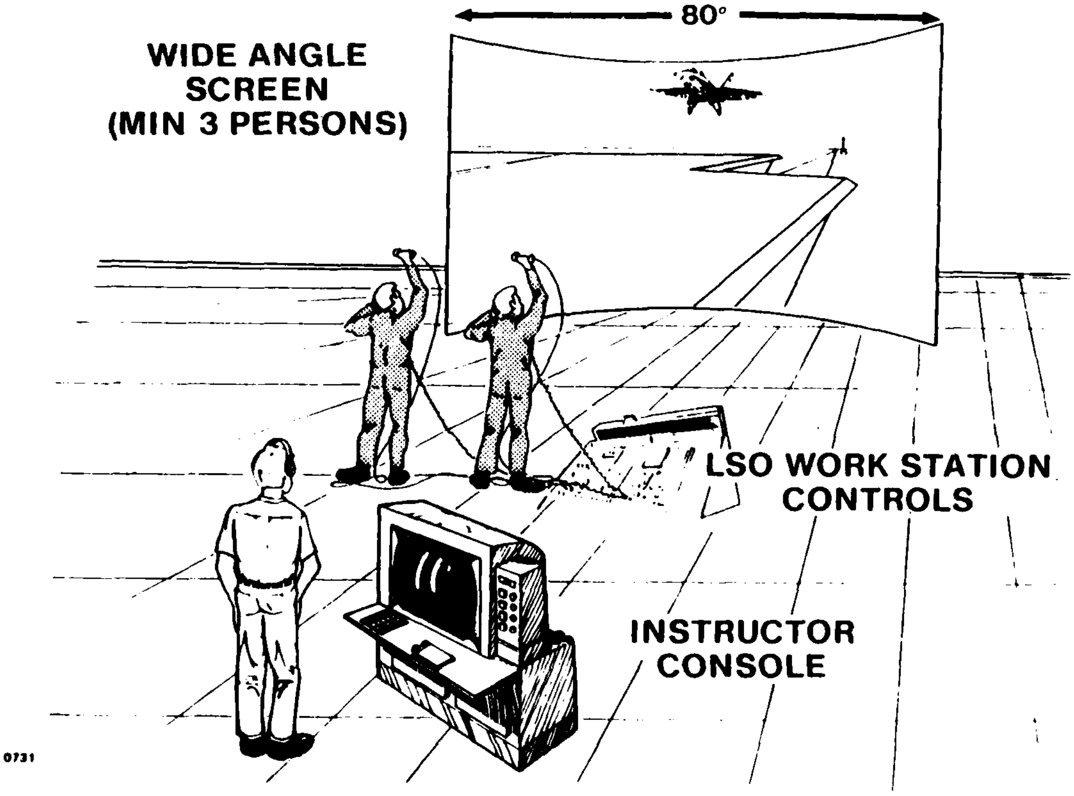


Figure 8. Characterization of proposed LSO Trainer System.

Another research project has explored the use of voice technology as the instructor's assistant. The overload on the instructor due to complex simulator based training generated the requirement for CVT to aid the instructor. Potential results include:

- Decrease in Instructor Requirements
- Decrease in Instructor Workload
- Consistency in Training
- Replay/Critique Capabilities
- Automated Interaction Among Instructor/Trainee/Simulator

CVT can also reduce manpower requirements in training by eliminating assistants and providing a wider range of applications in training.

## CVT IN TRAINING

Recent retention rate drops created shortages which negatively affect the training environment in terms of a lack of skilled personnel to perform as instructors and in interactive and team training. Worsening economic conditions which increase first time enlistments put an additional load on short staffed schools. Also, fleet shortages are increased due to slow and/or inadequate training. Additionally, Navy training and its related training devices are becoming increasingly complex and require increasingly sophisticated skills. Computer voice technology, in conjunction with other technologies in automated training and training device situations, can aid in the increase of training efficiency and can counteract training personnel shortages in these ways:

- simultaneous training
- personnel replacement
- instructor models
- interactive and team models
- trainee behavior models
- instructor assistance/support
- fidelity to actual systems

Instructor, interactive and trainee models are provided by both the PARTS and LSOTS using voice generation and recognition. PARTS provides a realistic environment under the supervision of an automated instructor. It also provides interactive modeling by providing the trainee air traffic controller with a pilot, pattern controller and tower controller models, thus reducing manpower requirements by relieving other personnel from performing these roles. Another result could be a decrease in training time as, often, trainees have been required to perform these roles for one another. Training can also be improved since the computer can provide far more flexible and accurate models than can trainees. In addition, PARTS includes a model of trainee behavior which provides quick and accurate performance measurement, and remedial and enrichment instruction as needed. Computer voice technology also provides fidelity in trainee data input and interactive personnel.

The LSOTS uses CVT to provide two-way communication between the pilot model and the LSO trainee. The LSOTS' instructor model could provide safe practice of critical situations with automated voice critique from the computer. LSO training had primarily an OJT training requirement; the LSOTS can reduce training time and personnel by greatly reducing the OJT requirement and could increase training safety.

The use of CVT as an instructor's assistant would reduce instructor busywork and increase instructor training effectiveness by facilitating man-machine interaction by both instructor and trainee. Presently, instructors often fail to utilize the full potential of a training device due to a number of reasons including:

- Training devices are becoming more and more complex.
- The high instructor turnover rate and the fact that training required for an instructor to become fully aware of a device's idiosyncrasies are often not provided for replacement instructors nor continued past the initial training and acceptance period.
- Instructor busywork chores (note taking, grading, and equipment monitoring) can create a high workload so that not all tasks are completed efficiently.

- The number of switches, lights, and displays often require an assistant operator for trainer utilization.

Computer voice technology can provide an effective man-machine communication channel. Coupled with a "resident instructor" in the form of a computer model, the human instructor would have access to a wealth of "knowledge" and assistance by having the capability to talk or converse with his training system -- as though it were an assistant. Automation can provide more efficient use of the device's potential, of the instructor's time, and thereby, increase training standardization.

Training payoff with this system results from increased training proficiency due to increased training device utilization and instructor effectiveness. Further, a reduction in the number of instructors can be realized from the capability of the trainee to interact directly with the instructor's assistant via computer voice technology. Thus, fewer human instructors can handle more trainees since many of the chores could be handled by the automated assistant. The trainee can make direct requests to the system for information, advice, or assistance via the voice channel, while the human instructor is free to concentrate on those trainees requiring more detailed assistance.

The use of CVT in an operational system, such as the projected application in the F-18 aircraft, will result in a requirement for CVT in the training device to provide fidelity to the actual system for F-18 pilot training.

There are two conditions which facilitate the application of CVT to military training. First, a high percentage of training devices are computerized, a necessary condition for CVT. Second, many non-training device applications would be trained with computer-based instruction. In this case the use of CVT would not be driven by a task speech requirement but by a need for a data entry method other than keyboard or light pen. This could be due to environmental effects or to eliminate interference with the program. Another non-training device use of CVT occurs when the required information cannot be presented by print due to low reading levels or environmental conditions, such as weather, poor light, or motion, which eliminate other presentation methods.

## Advantages Of CVT

The advantages of computer voice technology are broad-ranging and can provide fidelity and flexibility to computer use. CVT can model instructors, reducing instructor workloads and personnel requirements. CVT can provide more realistic training environments and increase student interest by providing models of required personnel such as team members. Data indicate that the prototype systems evaluated to date train no worse than traditional training. Voice technology has been shown to be particularly advantageous to industry when one or more of the following conditions apply:

- The worker's hands are busy.
- Mobility is required during the data entry process.
- The worker's eyes must remain fixed upon a display, an optical instrument, or some object to be tracked.
- The environment is too harsh to allow use of a keyboard.[18]

For example, a pilot needs his or her hands and eyes for actual flight. Voice technology can free the pilot from manual data entry and minimize the number of gauges to be viewed. In other applications, persons can directly access computer data by phone with CVT. Voice recognition can also improve the speed and accuracy of entering data into a computer. Many available voice components are relatively easy to interface with recommended computer hardware and software. Programming for voice recognition can usually be done in standard simple languages, such as Fortran and BASIC.

## Disadvantages Of CVT

The primary disadvantages of CVT are, at this time, cost and user acceptance. Other disadvantages are more accurately classified as subjects of current research, as discussed below:

- Speaker dependency is a major recognition disadvantage. Research in phonology and acoustics is being conducted to reduce or eliminate speaker dependency in a cost-effective manner.
- Language constraints such as non- and misrecognition negatively affect the user and his performance. Research geared towards establishing larger, more flexible and easily recognizable vocabularies will reduce language constraints.

- The software needed to operate voice systems must be developed for each application. Expanded use of CVT snould provide off-the-shelf software. The integration of voice capabilities into existing and planned operational and training systems presents interface complexities and requires refinement.

- Another disadvantage is the need to train and retrain the computer to accept individual voice patterns. Research to individualize for each talker a set of general patterns could speed this process.

- There may also be a requirement to train the individual to somewhat unnatural speech patterns. Research in different methods of identifying utterances can eliminate isolated word recognizers and the need for unnatural speech.

- Currently, there is no validated 'cookbook' for resolving the specific man-machine interface requirements. The Naval Training Equipment Center has, however, sponsored the development of an engineering guide.[19]

## Other Technology and CVT

Research in the CVT field can, and in many cases, has included interfacing computer voice technology with other technologies such as video disc, large screen displays, video gaming and artificial intelligence. Students can have voice response to, and control over, video disc systems during instruction. This mode of response entry is preferable to keyboard or light pen, especially when the student's hands are involved in other tasks.

Voice recognition can be used in instructional situations to control large screen displays. One example is of a cockpit trainer using a one-hundred and-eighty-degree computer-generated-imagery large screen display. The pilot-trainee communicates with the computer operator who then controls the display. Voice recognition could decrease the training time spent, as the display would be changed more rapidly and would release the computer operator's time. Another example is an air traffic control trainer with a one-hundred-and-twenty-degree display using slides and 16mm projectors. This trainer has four computer terminals requiring five instructor-operators and two or three more instructors in the "tower" to train three students. The instructor-operators act as pilots and program their aircrafts' movements across the display. Voice recognition and speech generation could replace the five instructor-operators in this case.

Students can have voice interaction with trainers using video gaming techniques, such as in the Combat Engineer Vehicle Trainer. In this trainer the student would vocally command team members and vehicle actions which he views on the video display. Voice here provides greater fidelity to actual command situations.

Currently, computer models of the listener constitute attempts to duplicate reasoning processes. This is done by providing the computer with some form of higher level knowledge sources which may be considered synonymous with artificial intelligence. A computer system that incorporates the factual knowledge of linguistics and of the language plus the heuristic knowledge of a human listener would constitute an artificial intelligence system capable of achieving a high degree of voice recognition accuracy as well as speech understanding.[20]

## ISD

The Interservice Procedures for Instructional Systems Development, or ISD Model, has been modified by the various services to suit their respective needs. The Air Force has expanded the model to incorporate training devices, and the Navy has extended the model to include speech system design. Since training device design and development is largely an engineering task, the primary involvement of the instructional developer is in analysis and pre-engineering design phases. It is in these phases where the initial device design concept is determined based on an analysis of the tasks. The device design concept includes the functions and features required for the device to instruct the learner to perform the tasks. This concept is then turned over to the engineers who develop the design specification for the device.

The ISD model notes five phases in instructional development - analysis, design, development, implementation, and control. The analysis phase consists of, among other things, job analysis, task selection, job performance measure construction, existing course analysis, and instructional setting selection. When analyzing a job to determine job requirements, the developer may cycle through these steps several times. This cyclic analysis is critical in determining the features of a complex training device. The determination of how to satisfy certain training needs will depend on the state of the art, costs, and the integration complexities of various technologies.

A needs analysis identifies a discrepancy between what is and what ought to be. Should it be determined that the discrepancy requires a training solution, the ISD model would be utilized. First, the job is defined and broken into validated tasks with conditions, cues, standards and elements. Tasks are then selected for training using performance, criticality, and timing data. In an environment where a training device is not a consideration, the developer determines job performance requirements, reviews existing training courses, determines the instructional setting, and moves on to the design phase. Although each of these steps is complex, the addition of training devices further complicates the process.

If a device is a consideration, this is determined at the task selection stage. Once the training tasks have been selected, they must be divided into those which require hands-on training and those which do not. The process requires looking at the skills and knowledge related to the task elements. The traiuing device tasks must then be analyzed to determine the instructional features needed on the device to effectively train the student. Instructional features are those features of the trainer which are involved in training. The selection of instructional features is dependent on the nature of the trainer - maintenance, operator, scenario, or part-task - on the performance environment and on the required fidelity. Features include cues, feedback, performance measurement, malfunctions, dials, indicators, scopes, screens, time, controls, speech, motion, visuals, audio, data inputs and outputs, and the manipulation of the features themselves. The selected features must be looked at in relationship to one another, and within the constraints of the program. This means consideration of integration, costs, time, and technology state-of-the-art. A realistic preliminary device description can then be made to be turned over to the engineers.

## CVT In ISD

More specifically, the determination of a CVT requirement can be shown as a subroutine of the instructional development model. The Navy sponsored the development of an engineering design guide for CVT which forms the basis for this subroutine and details voice system design with two exceptions.[21] This guide does not specify the factors which signal a speech requirement nor the integration of CVT with other computerized technologies.

When selecting tasks which can be trained using training devices, one of the selection criteria is a speech requirement. The following situations indicate the need for a training device with a speech generation and/or voice recognition capability:

- In tasks where CVT is used in the operational or maintenance environment. In systems where voice is used to control the computer and/or to issue information from the computer, CVT is required for training.

- In speech related tasks, where vocal advisories or commands are required, or where there is voice interaction. Included are such tasks as air traffic control, air intercept control, and landing signal officer.

- For data entry in computerized training systems when the hands and eyes are otherwise occupied, as in air traffic control and vehicle or aircraft operation training.

- Where voice is a modality for measuring performance. Voice recognition can be used to measure a student's performance in primarily vocal tasks such as air traffic control, and command and control.

- To stimulate verbal communications. Speech generation can ask questions or give directions which require verbal responses from the student.

- When voice feedback is required. Speech generation can provide fast and natural vocal feedback as needed.

- When the instructor has a heavy workload in equipment set-up, CVT can aid instructors in dealing with complex trainers or other non-instructional burdens. CVT can provide a computerized instructor's assistant to aid in teaching and/or provide administrative assistance.

- When the training task requires high instructor interaction with the student, CVT can provide an instructor model and reduce the manpower required for teaching certain topics.

- For team training when all members of a team cannot be trained at once. This eliminates the need for additional manpower for role-playing. This also allows stricter control of the instructional situation by allowing control over and variance of the performance of the modeled team members.

- For heads-up, hands-busy situations such as maintenance training. When the student's hands are busy with the maintenance task, he can control the instruction vocally. The student can request the next frame or ask for a repeat of a frame. If his eyes are also busy, speech generation can direct the task. Voice recognition can also be employed where keyboard operation by the student is not desired, and light pen is not an acceptable solution.

In addition, both performance measurement and environment issues must be considered in determining a speech requirement.

Once the speech tasks have been selected, the determination of the nature and feasibility of CVT as a speech alternative can begin. This step would be concurrent with the same procedures for other computerized technologies. In fact, the feasibility determination will also require consideration of all computer technologies as a unit in order to assess integration problems and associated hardware and software costs.

Since the Navy CVT design guides are for engineers and consequently, very detailed, the developer should perform a cursory analysis to determine the feasibility of using CVT. If feasible, the developer and/or the engineer can then perform the more detailed analysis. A computer speech technology specialist should be involved during this process. These are the steps in CVT system design:

1. Establish Vocabulary

2. Identify Voice Technology System Design Requirements

3. Determine Voice Technology State-of-the-Art

4. Project Voice Technology Capability

5. Make Design Decisions

6. Develop Operating and Human Factors Design

7. Develop Voice System Design Requirements Specifications

These steps are the same for both speech generation and voice recognition, although the procedures for steps two, three and five differ. Steps one, four, six and seven should be carried out concurrently when considering both CVR and CSG. For our purpose, voice recognition will be considered first.

A surface analysis of CVR requires consideration of the following factors:

- Isolated vs. Connected Recognition

- Vocabulary size

- Speaker Dependency

- Task Criticality

- Voice Collection

- Environment

These factors must then be related to the state-of-the-art to determine if the technology exists and if so, related costs. If the system seems feasible, then the more detailed analysis must be performed, resulting in a speech system requirements analysis.

Now let's consider design of computer speech generation. The same general procedures apply.

A cursory analysis of CSG should look at the required vocabulary size, the number of speaker voices, the required voice quality and the complexity of transmissions. These can be generally related to the state-of-the-art to determine if the technology exists and if so, the related costs. If the system seems feasible, then a more detailed analysis must be performed resulting in a speech system requirements specification.

Once the voice system has been determined to be feasible, the process returns to the training device development model where making design decisions and developing operating and human factors designs actually occur in conjunction with the same analysis as for the device. These functions are particularly critical for highly complex technological devices, as the developer must also consider the integration of technologies. A system may also require a highly complex and flexible visual system such as utilizing video discs. For example, both the hardware integration and software design issues must be considered before finalizing the system specification for either the video disc or speech systems. In addition, technology advances must be determined and planning completed prior to matching the release of these advances to the device production schedule.

Concurrent with the design of the training device is the design of the rest of the training system. The ISD model design phase determines objectives, tests, entry behavior, and the sequence and structure of the training to include use of the training device. The rest of the ISD process - development, implementation and control -

should occur concurrently and interactively with the actual development of the training device. Course development must accommodate the training at all times so that the resulting training is integrated, effective, and efficient.

## Human Factors

In designing instruction using CVT, there are many human factors issues to consider.[22] The primary issues are as follows:

- Validation - machine training, retraining, modeling of voice technology
- User frustration, stress, fatigue and boredom
- User task training
- User system acceptance
- Environment

In speaker dependent systems, the speaker must input from one to ten samples of each utterance depending upon the hardware requirements of the system selected, so that the system can recognize the user's words. The newly-collected utterances should be validated before continuing. With a large vocabulary, training the computer can be a tedious task for the user. This redundancy can be minimized by having the user train utterances as needed throughout the actual task training or by selecting a device requiring few voice samples, although such devices are usually more expensive. In addition, either the user or the system must be able to recognize when retraining may be needed, primarily in situations of non- and misrecognition. If the system is being used over an extended period of time, the user may want to enter one sample of each word every day or at each use. The system should be designed to cue the user to retrain. This type of trainee feedback leads to internalization by the student of the concepts of how speech is recognized by computers.

Misrecognition, unnatural speech patterns, machine training, the environment, and the task itself can all lead to user frustration, stress, fatigue and boredom. All of these can, in turn, effect the user's voice and cause recognition difficulties. Therefore, the designer must take care to convey to the student the significance of these factors. Then, when the student can recognize these factors, internalization is beginning.

One of the most important human factors issues is user system acceptance. In CSG, the user must be able to relate to the speech quality of the computer. In a simple, non-critical task a robotic sounding voice may be acceptable and perhaps, even interesting. But in a complex, critical task, such as AIC training, robotic disjointed speech can be distracting and frustrating when the user needs to hear natural-sounding responses from the pilot model. Therefore, the user may need to listen to the computerized speech before actual task training begins in order to familiarize himself with its sound and to minimize any distraction. In phoneme generated speech, familiarization can increase the user's level of comprehension. In CVR the user must be able to relate comfortably to speaking to a machine. The extent of the user training required depends on the system used. Isolated word recognition systems require significant pauses between words or phrases which may be unnatural for the speaking requirement. Easily confused words, if they must be used, may require using a different verbal word, such as "execute" for "run" which the user must remember. The user must also try to keep his voice consistent to achieve the best recognition. In CVR the user must also be trained to use and train the system itself. The user may also either over- or underestimate the capabilities of a CVT system. The user may see it as a "Star Wars" system which is intelligent and can understand anything that is said, then become frustrated when it doesn't react like a human. On the other hand, the user may not believe the system's capabilities and may not use it to its full extent. Both of these situations can be handled with user training in CVT, especially, if the training incorporates a "user friendly" design which will be discussed later.

The environment must also be considered in using CVT. A noisy environment, such as machinery or aircraft noise, may require repeated or delayed inputs. Another noise problem can result from the user in terms of coughing, sneezing, or throat clearing. The computer will attempt to recognize these sounds as words. Certain types of earphones or microphones may be required to help minimize noise problems.

## CVT Design

Once the technical content of the training is organized including the use of CVT, the CVT portions can be designed for optimal use and acceptance. At this stage the system hardware has been defined including microphones and earphones. The instructional developer must then design the instruction for the software developer. The instructional developer has all of the usual considerations - cues, feedback, prompting, exercises, review, branching, and so on. For CVT he must also determine how and when to train and retrain vocabulary, model the instructor, design CVT feedback and cues, and provide user training.

The use of connected or isolated recognition should have been determined by the design phase. The type of voice input required must be incorporated into the design. The user must also be trained to pause between words or phrases or to structure his inputs in certain formats.

The user must also be trained to train the machine with the necessary vocabulary. The design should incorporate vocabulary training as needed throughout the actual task training. The user should also receive some training in CVT as needed during actual task training. If the user has a special requirement or if the training task is particularly complex, then the developer may want to consider a job aid. This technique can be accomplished either via a handout or via software, depending upon cost consideration. The job aid would include such information as training, retraining and pausing. User training should include training the system, recognizing unnatural voice patterns, pausing, and becoming familiar with vocal noises, such as coughing, sneezing, and throat clearing.

The design of the task training must take into consideration the constraints of the technologies being used to support the training. Complex task training could be beyond the capabilities of the state-of-the-art of the technologies or could require highly complex, expensive programming and interfaces.

Both the user training and the task training can be made very acceptable to the user if the system can accommodate computer feedback either visually or by computer speech generation. The computer can tell the user where his recognition problems lie. For example:

1. The computer can list several words which were closest to the misrecognized word and give a score for each possible choice. The score might be computed from one to 10, with 10 being the most recognizable. For example, if the misrecognized word is 'four,' the computer can say, "What you said sounded like 'door' - 8.3, 'pour' - 8.1, 'four' - 7.8, and 'floor' - 7.4." This helps the user identify areas of mismatch and encourages him to retrain. The computer, in fact, might recommend retraining within preset parameters. It also helps the user develop a model of what a computer 'ear' hears when a human speaks.

2. The computer can determine if misrecognition is due to too loud or too soft speech and tell the user. If the user finds this volume to be normal, he may wish to retrain the computer.

3. The computer can be programmed to reject or not recognize any word which falls below a certain score or acceptance threshhold. The computer can say "What you said was scored 4.1, which is below the acceptance threshhold of 6.5."

The feedback can be very informal and chatty if this is acceptable, although this could be expensive in terms of additional software.

The vocabulary size and type of system will, to a large part, determine the design. If a task requires a small vocabulary and connected speech recognition, these requirements could be trained prior to the onset of actual task training. However, training the computer as needed during the task (in-context) is probably the most acceptable to the user.

Points to consider when designing computer voice technology training follow:

- Encourage warm-up, relaxing, breaks, practice, and retraining.
- Emphasize consistency.
- Prompt the user in saying the words. If CSG is used, this can be done vocally.
- Allow for as much in-context training as possible.
- Introduce the vocabulary as needed.
- Keep repetition to a minimum.
- Validate newly collected words or phrases before proceeding. Retrain if a misrecognition is detected.
- Update voice samples during the training procedures.
- Discourage extraneous noise such as coughing, sneezing, and throat clearing.
- Tradeoff between non- and misrecognition through softwaro control.

Computer voice technology is a viable training component which is rapidly growing in its ability to provide flexibility, fidelity and sophistication to the training environment. The instructional designer can maintain control over the use of new technologies such as CVT in training by achieving an awareness and facility as offered by this paper.

# REFERENCES

1. Lea, Wayne A. Computer Recognition of Speech, Seminar Workbook for a Short Course, Student Edition. Santa Barbara, CA: Speech Science Publications, 1980, Chapter 11.

2. Integrated Computer Systems, Voice Input/Output: The State of the Art, 1980, page II-2-3.

3. Naval Training Equipment Center, Interim Report - Computer Speech Technology's R&D Plan Development, 1980, page 7.

4. Voice Input/Output, page II-1-2.

5. Voice Input/Output, page II-2-3.

6. Interim Report, page 9.

7. Baker, Janet MacIver. Brief Status Summary for Automatic Speech Recognition at the Start of the 80's, SAE Technical Paper Series, No. 800195. Warrendale, PA: Society of Automotive Engineers, Inc., 1980.

8. Interim Report, page 9.

9. Interim Report, page 7.

10. Interim Report, page 16.

11. Chatfield, Douglas C., Marshall, Philip H., and Gidcumb, Charles F. Naval Training Equipment Center, Instructor Model Characteristics for Automated Speech Technology (IMCAST), 1979. Technical Report NAVTRAEQUIPCEN 79-C-0085-1.

12. Hicklin, Mary, et al. Naval Training Equipment Center, Ground Controlled Approach Controller Training System, Final Technical Report, 1980. Technical Report NAVTRAEQUIPCEN 77-C0-0162-6, page 6-10.

13. McCauley, M. E., and Semple, C. A. Naval Training Equipment Center, Precision Approach Radar Training System (PARTS) Training Effectiveness Evaluation, 1980. Technical Report NAVTRAEQUIPCEN 79-C-0042-1.

14. Grady, Michael W., Porter, J. E., Satzer, Jr., W. J., and Sprouse, B. D. Naval Training Equipment Center, Speech Understanding in Air Intercept Controller Training System Design, 1978. Technical Report NAVTRAEQUIPCEN 78-C-0044-1, pages 7-8.

15. McCauley, Michael E., Root, Robert W., and Muckler, Frederick A. Naval Training Equipment Center, Training Evaluation of an Automated Training System for Air Intercept Controllers, 1982. Technical Report NAVTRAEQUIPCEN 81-C-0055-1.

16. Chatfield, pages 13-17.

17. Popelka, Beverly A., and Knerr, C. Mazie. Defense Advanced Research Projects Agency, Team Training Applications of Voice Processing Technology, 1980. Contract No. MDA903-79-C-0209 ARPA, pages 18-23.

18. Voice Input/Output, page II-1-3.

19. Cotton, John C., and McCauley, Michael E. Naval Training Equipment Center, Voice Technology Design Guide for Navy Training System, in press. Technical Report NAVTRAEQUIPCEN 80-C-0057-1.

20. Interim Report, page 16.

21. Voice Technclogy Design Guide.

22. Van Hemel, Paul E., Van Hemel, Susan B., King, William J., and Breaux, R. Naval Training Equipment Center, Training Implication of Airborne Applications of Automated Speech Recognition Technology, 1980. Technical Report NAVTRAEQUIPCEN 80-C-0009-0155-1.

## REPORT DOCUMENTATION PAGE

| | | | |
|---|---|---|---|
| 1. Recipient's Reference | 2. Originator's Reference | 3. Further Reference | 4. Security Classification of Document |
| | AGARD-LS-129 | ISBN 92-835-0331-7 | UNCLASSIFIED |

5. Originator
Advisory Group for Aerospace Research and Development
North Atlantic Treaty Organization
7 rue Ancelle, 92200 Neuilly sur Seine, France

6. Title

SPEECH PROCESSING

7. Presented at a Lecture Series under the sponsorship of the Avionics Panel and the Consultant and Exchange Programme of AGARD on 20–21 June, 1983 in Trondheim, Norway; on 23–24 June, 1983 in Copenhagen, Denmark and on 27–28 June, 1983 in Delft, The Netherlands.

| | |
|---|---|
| 8. Author(s)/Editor(s)<br>Various | 9. Date<br>May 1983 |
| 10. Author's/Editor's Address<br>Various | 11. Pages<br>126 |

12. Distribution Statement
This document is distributed in accordance with AGARD policies and regulations, which are outlined on the Outside Back Covers of all AGARD publications.

13. Keywords/Descriptors

Speech recognition
Speech analysis
Voice communication
Avionics
Data processing

14. Abstract

Lecture Series 129 is concerned with speech processing and is sponsored by the Avionics Panel of AGARD and implemented by the Consultant and Exchange Programme.

The aim of the lectures is to familiarize the participants with the potential applications of speech processing (and in particular the military applications). The Lecture Series presents the state-of-the-art in the areas of research in speech recognition, isolated word recognition systems, automatic speaker identification, test and evaluation of automatic word recognition systems, and it covers applications of speech processing to avionics.

AGARD Lecture Series No.129
Advisory Group for Aerospace Research and Development, NATO
SPEECH PROCESSING
Published May 1983
126 pages

Lecture Series 129 is concerned with speech processing and is sponsored by the Avionics Panel of AGARD and implemented by the Consultant and Exchange Programme.

The aim of the lectures is to familiarize the participants with the potential applications of speech processing (and in particular the military applications). The Lecture Series presents the state-of-the-art in the areas

P.T.O

AGARD-LS-129

Speech recognition
Speech analysis
Voice communication
Avionics
Data processing

AGARD Lecture Series No.129
Advisory Group for Aerospace Research and Development, NATO
SPEECH PROCESSING
Published May 1983
126 pages

Lecture Series 129 is concerned with speech processing and is sponsored by the Avionics Panel of AGARD and implemented by the Consultant and Exchange Programme.

The aim of the lectures is to familiarize the participants with the potential applications of speech processing (and in particular the military applications). The Lecture Series presents the state-of-the-art in the areas

P.T.O

AGARD-LS-129

Speech Recognition
Speech analysis
Voice communication
Avionics
Data processing

AGARD Lecture Series No.129
Advisory Group for Aerospace Research and Development, NATO
SPEECH PROCESSING
Published May 1983
126 pages

Lecture Series 129 is concerned with speech processing and is sponsored by the Avionics Panel of AGARD and implemented by the Consultant and Exchange Programme.

The aim of the lectures is to familiarize the participants with the potential applications of speech processing (and in particular the military applications). The Lecture Series presents the state-of-the-art in the areas

P.T.O

AGARD-LS-129

Speech recognition
Speech analysis
Voice communication
Avionics
Data processing

AGARD Lecture Series No.129
Advisory Group for Aerospace Research and Development, NATO
SPEECH PROCESSING
Published May 1983
126 pages

Lecture Series 129 is concerned with speech processing and is sponsored by the Avionics Panel of AGARD and implemented by the Consultant and Exchange Programme.

The aim of the lectures is to familiarize the participants with the potential applications of speech processing (and in particular the military applications). The Lecture Series presents the state-of-the-art in the areas

P.T.O

AGARD-LS-129

Speech recognition
Speech analysis
Voice communication
Avionics
Data processing

of research in speech recognition, isolated word recognition systems, automatic speaker identification, test and evaluation of automatic word recognition systems, and it covers applications of speech processing to avionics.

The material in this publication was assembled to support a Lecture Series under the sponsorship of the Avionics Panel and the Consultant and Exchange Programme of AGARD presented on 20–21 June 1983 in Trondheim, Norway; on 23–24 June 1983 in Copenhagen, Denmark and on 27–28 June 1983 in Delft, The Netherlands.

ISBN 92-835-0331-7

of research in speech recognition, isolated word recognition systems, automatic speaker identification, test and evaluation of automatic word recognition systems, and it covers applications of speech processing to avionics.

The material in this publication was assembled to support a Lecture Series under the sponsorship of the Avionics Panel and the Consultant and Exchange Programme of AGARD presented on 20–21 June 1983 in Trondheim, Norway; on 23–24 June 1983 in Copenhagen, Denmark and on 27–28 June 1983 in Delft, The Netherlands.

ISBN 92-835-0331-7

of research in speech recognition, isolated word recognition systems, automatic speaker identification, test and evaluation of automatic word recognition systems, and it covers applications of speech processing to avionics.

The material in this publication was assembled to support a Lecture Series under the sponsorship of the Avionics Panel and the Consultant and Exchange Programme of AGARD presented on 20–21 June 1983 in Trondheim, Norway; on 23–24 June 1983 in Copenhagen, Denmark and on 27–28 June 1983 in Delft, The Netherlands.

ISBN 92-835-0331-7

of research in speech recognition, isolated word recognition systems, automatic speaker identification, test and evaluation of automatic word recognition systems, and it covers applications of speech processing to avionics.

The material in this publication was assembled to support a Lecture Series under the sponsorship of the Avionics Panel and the Consultant and Exchange Programme of AGARD presented on 20–21 June 1983 in Trondheim, Norway; on 23–24 June 1983 in Copenhagen, Denmark and on 27–28 June 1983 in Delft, The Netherlands.

ISBN 92-835-0331-7