NORTH ATLANTIC TREATY
ORGANISATION

RESEARCH AND TECHNOLOGY
ORGANISATION

AC/323(IST-037)TP/22

www.rta.nato.int

RTO TECHNICAL REPORT **TR-IST-037**

Use of Speech and Language Technology in Military Environments

(La mise en œuvre des technologies de la parole
et du langage dans les environnements militaires)

The material in this publication was assembled to support a Lecture Series under the sponsorship of the Information Systems Technology Panel (IST) presented on 20-21 November 2003 in Montreal, Canada; 1-2 December 2003 in Arcueil, France; and 4-5 December in Istanbul, Turkey.

.

Published December 2005

Distribution and Availability on Back Cover

NORTH ATLANTIC TREATY
ORGANISATION

RESEARCH AND TECHNOLOGY
ORGANISATION

AC/323(IST-037)TP/22

www.rta.nato.int

RTO TECHNICAL REPORT **TR-IST-037**

Use of Speech and Language Technology in Military Environments

(La mise en œuvre des technologies de la parole
et du langage dans les environnements militaires)

by

Dr. Stéphane PIGEON, Belgium

Mr. Carl SWAIL (Secretary), Canada

Dr. Edouard GEOFFROIS, France

Ms. Christine BRUCKNER, Germany

Dr. David van LEEUWEN, The Netherlands

Prof. Carlos TEIXEIRA, Portugal

Mr. Ozgur ORMAN, Turkey

Mr. Paul COLLINS, United Kingdom

Dr. Timothy ANDERSON (Chairman), USA

Mr. John GRIECO, USA

Dr. Marc ZISSMAN, USA

The material in this publication was assembled to support a Lecture Series under the sponsorship of the Information Systems Technology Panel (IST) presented on 20-21 November 2003 in Montreal, Canada; 1-2 December 2003 in Arcueil, France; and 4-5 December in Istanbul, Turkey.

The Research and Technology Organisation (RTO) of NATO

RTO is the single focus in NATO for Defence Research and Technology activities. Its mission is to conduct and promote co-operative research and information exchange. The objective is to support the development and effective use of national defence research and technology and to meet the military needs of the Alliance, to maintain a technological lead, and to provide advice to NATO and national decision makers. The RTO performs its mission with the support of an extensive network of national experts. It also ensures effective co-ordination with other NATO bodies involved in R&T activities.

RTO reports both to the Military Committee of NATO and to the Conference of National Armament Directors. It comprises a Research and Technology Board (RTB) as the highest level of national representation and the Research and Technology Agency (RTA), a dedicated staff with its headquarters in Neuilly, near Paris, France. In order to facilitate contacts with the military users and other NATO activities, a small part of the RTA staff is located in NATO Headquarters in Brussels. The Brussels staff also co-ordinates RTO's co-operation with nations in Middle and Eastern Europe, to which RTO attaches particular importance especially as working together in the field of research is one of the more promising areas of co-operation.

The total spectrum of R&T activities is covered by the following 7 bodies:

- AVT Applied Vehicle Technology Panel
- HFM Human Factors and Medicine Panel
- IST Information Systems Technology Panel
- NMSG NATO Modelling and Simulation Group
- SAS Studies, Analysis and Simulation Panel
- SCI Systems Concepts and Integration Panel
- SET Sensors and Electronics Technology Panel

These bodies are made up of national representatives as well as generally recognised 'world class' scientists. They also provide a communication link to military users and other NATO bodies. RTO's scientific and technological work is carried out by Technical Teams, created for specific activities and with a specific duration. Such Technical Teams can organise workshops, symposia, field trials, lecture series and training courses. An important function of these Technical Teams is to ensure the continuity of the expert networks.

RTO builds upon earlier co-operation in defence research and technology as set-up under the Advisory Group for Aerospace Research and Development (AGARD) and the Defence Research Group (DRG). AGARD and the DRG share common roots in that they were both established at the initiative of Dr Theodore von Kármán, a leading aerospace scientist, who early on recognised the importance of scientific support for the Allied Armed Forces. RTO is capitalising on these common roots in order to provide the Alliance and the NATO nations with a strong scientific and technological basis that will guarantee a solid base for the future.

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

Published December 2005

ISBN 92-837-1144-0

Table of Contents

Page

List of Figures vi

List of Tables vii

Preface viii

Information System Technology Research Task Group 001 ix

Executive Summary and Synthèse ES-1

Chapter 1 – INTRODUCTION 1-1

Chapter 2 – SPEECH AND LANGUAGE IN MILITARY APPLICATIONS 2-1

2.1 Command and Control 2-2

2.2 Communications 2-4

2.3 Computers and Information Access 2-5

2.4 Intelligence 2-6

2.5 Training 2-7

2.5.1 Language Training 2-7

2.5.2 Simulation 2-7

2.5.3 Requirements 2-7

2.6 Multinational Forces 2-7

Chapter 3 – AVAILABLE TECHNOLOGIES 3-1

3.1 Speech Processing 3-1

3.1.1 Speech Coding 3-1

3.1.1.1 Current NATO Standards 3-3

3.1.1.2 New NATO Standard on Speech Coding 3-3

3.1.2 Speech Enhancement 3-4

3.1.2.1 Enhancement Techniques 3-5

3.1.2.2 Speech Under "Stress" Technology 3-11

3.1.3 Stress Reduction Techniques 3-12

3.1.4 Speech Analysis 3-14

3.1.4.1 Voice Stress Analysis 3-15

3.1.5 Speech Synthesis 3-17

3.1.5.1 Text-to-Speech = Linguistic Processing + Speech Generation 3-17

3.1.5.2 Linguistic Processing 3-17

3.1.5.3 Speech Generation 3-18

3.1.5.4 Rule-Based versus Concatenative Approaches 3-19

3.1.5.5 State of the Art 3-19

3.1.5.6 Military Applications 3-20

3.1.6 Speech Recognition 3-20
3.1.6.1 Speaker Dependent or Independent Recognition 3-21
3.1.6.2 Vocabulary Size 3-21
3.1.6.3 Large Vocabulary Continuous Speech Recognition 3-22
3.1.6.4 Robust Recognition for Adverse Conditions 3-23
3.1.7 Speaker Recognition 3-25
3.1.7.1 Speaker Verification 3-25
3.1.7.2 Speaker Detection 3-26
3.1.7.3 Other Applications 3-26
3.1.8 Language Identification 3-27
3.1.9 Accent Identification 3-27
3.2 Language Processing 3-29
3.2.1 Topic Spotting 3-29
3.2.2 Translation 3-30
3.2.3 Understanding 3-32
3.3 Interaction 3-32
3.3.1 Interactive Dialogue 3-32
3.3.2 Multi-Modal Communication 3-33
3.3.3 Dimensional Sound Display 3-34
3.4 References 3-36

Chapter 4 – ASSESMENT AND EVALUATION 4-1

4.1 General Principles of Specification and Assessment of Speech and Language Processing Systems 4-1
4.1.1 Introduction 4-1
4.1.2 The Right Application using the Right Technology 4-1
4.1.3 System Specification 4-2
4.1.4 Evaluation and Assessment 4-3
4.1.5 Standards and Resources 4-4
4.1.6 Corpora 4-4
4.1.7 Other Resources 4-5
4.2 Specification and Assessment of Speech Communication Systems 4-5
4.2.1 Introduction 4-5
4.2.2 Intelligibility Measures and Quality Rating 4-6
4.2.2.1 Tests at Phoneme and Word Level 4-6
4.2.2.2 Tests at Sentence Level 4-7
4.2.2.3 Quality Rating 4-7
4.2.2.4 Objective Intelligibility Measures 4-7
4.2.2.5 Criteria and Relation between Various Measures 4-7
4.2.3 Speech Communicability Testing 4-8
4.3 Specification and Assessment of Speech Recognition Systems 4-11
4.3.1 Introduction 4-11
4.3.2 Recognition Systems 4-11
4.3.2.1 Word Recognition 4-11
4.3.2.2 Speaker Classification 4-12
4.4 References 4-14

Chapter 5 – CASE STUDIES AND (FUTURE) APPLICATIONS 5-1
5.1 Cockpit Fast Jet 5-1
5.2 Helicopter 5-2
5.3 Sonar 5-3
5.4 Noise Reduction 5-4
5.5 Training of Air Traffic Controllers 5-6
5.6 DARPA Spoken Language Systems Demonstrations and Applications 5-7
5.7 Battlefield Battle Management System 5-8
5.8 Speech Coders 600-1200 Bps 5-11
5.9 Conclusion 5-12

Appendix 1 – Glossary of Speech and Language Technology Terms A1

Appendix 2 – List of Authors on Original Report A2

Appendix 3 – List of Authors for Updated Report A3

List of Figures

Figure		Page
Figure 2.1	Control Loop of Human Decision Making	2-2
Figure 3.1	Attenuation Obtained by Conventional and ANR Headsets	3-6
Figure 3.2	STI as a Function of the Noise Level for Two Different Microphones and Two Speaking Distances	3-7
Figure 3.3	General Characteristics of Three Types of Automatic Level Control	3-8
Figure 3.4	Impulse Removal as a Function of Impulse Rate and Duty Cycle for a 1 millisec Impulse Width	3-9
Figure 3.5	The Effects of Stress on Automatic Speaker Recognition	3-13
Figure 3.6	The Effects of Stress on Automatic Speech Recognition and the Improvement Obtained Using an Equalization Process on a 30 Word Vocabulary	3-13
Figure 3.7	Pairwise Stress Classification using the Teager Energy Operator	3-14
Figure 3.8	Diogenes Lantern Output Indicating No Stress	3-15
Figure 3.9	Diogenes Lantern Output Indicating Voice Stress	3-16
Figure 3.10	The Truster Stress Analyzer as Displayed on a Monitor	3-16
Figure 3.11	Results of a Number Identification Experiment, in which a 3D Auditory Display was Compared with Normal Sound Presentation	3-35
Figure 4.1	A Model of the Relationship between the Applications of Speech and Language Systems and the Underpinning Technology	4-2
Figure 4.2	Relationship between some Intelligibility Measures according to Steeneken (1992) and DIS ISO 9921	4-8
Figure 4.3	Average Subject Responses Time on Part of the Task (representative for communication efficiency) for Four Different Levels of Background Noise (average speech level at normal vocal effort was 70 dB(A)).	4-9
Figure 4.4	Average Acceptability Rating for Four Different Levels of Background Noise (average speech level at normal vocal effort was 70 dB(A)).	4-10
Figure 4.5	Average Subject Responses Time on Part of the Task (representative for communication efficiency) for Three Different Roundtrip Delay Values	4-10
Figure 4.6	Receiver Operating Characteristic for Speaker Identification System	4-13
Figure 5.1	The SEU Noise Reduction System	5-6
Figure 5.2	The ALPS Display showing a Wide Angle IR Display	5-11

List of Tables

Table		Page
Table 1.1	Relationship between Military Applications of Speech and Language Processing and Available Technologies	1-2
Table 2.1	Functional Analysis of Two Command and Control Examples	2-3
Table 3.1	Contamination	3-5
Table 3.2	Enhancement Techniques	3-6
Table 3.3	Classes of Electronic Enhancement	3-9
Table 3.4	Examples of Stressors that Impact on the Production of Human Speech	3-12
Table 3.5	Table of Voice Stress Analysis Products	3-15
Table 3.6	Large Vocabulary Continuous Speech Recognition System Performance	3-22
Table 3.7	Mean Recognition Scores (mean % correct) and Standard Errors (se %) based on One Speaker, Ten Test Runs, Seven Noise Conditions and With and Without the Oxygen Mask	3-24
Table 3.8	Word Error Rates for Test of Canadian Marconi Company Recognizer in Bell 412 Helicopter in Three Flight Exercises	3-24
Table 3.9	Word Error Rates for Test of Verbex Speech Commander in Bell 412 Helicopter in Three Flight Conditions and a Quiet Hanger Test	3-24
Table 3.10	Speaker Verification Performance Specification	3-26
Table 3.11	Confusion Matrix (%) for an Accent Identification System	3-28
Table 5.1	CVC-Word Scores (m %) for Male and Female Speech based upon 16 Speaker-Listener Pairs	5-12

Preface

Efficient speech communication is recognized as a critical and instrumental capability in many fields of military activities such as command and control, aircraft, ship and ground vehicle operations, military communications, translation, intelligence, and training. The NATO Research Task Group on Speech and Language Technology (AC/323/IST-011/RTG-001) has, since its establishment as Research Study Group 10 (RSG.10) in 1978, conducted experiments and surveys focused on military applications of speech and language processing.

Guided by its mandate, the Research Task Group initiated in the past the publication of overviews on potential applications of speech technology for use in NATO forces (Beek et al. 1977, Weinstein 1991, Steeneken et al. 1996) and also organized workshops and lecture series on military-relevant speech technology topics.

In recent years, the speech R & D community has developed or enhanced a number of technologies which can now be integrated into a wide range of military applications and systems:

- Speech coding algorithms are used in very low bit-rate military voice communication systems. These state-of-the-art coding systems increase the resistance against jamming;
- Speech input and output systems can be used in command and control environments to substantially reduce the workload of operators. In many situations operators have busy eyes and hands, and must use other means to trigger control functions and receive feedback information;
- Large vocabulary speech recognition and speech understanding systems are useful as training aids and to prepare for missions;
- Speech processing techniques are available to identify talkers, languages, and keywords and can be integrated into military intelligence systems;
- Computer-assisted training systems combining automatic speech recognition and synthesis technologies can be utilized to train personnel (e.g. air traffic controllers) with minimum or no instructor support.

This updated report reviews the wide range of potential military applications and also describes the state-of-the-art in speech technology. The IST-011/RTG-001 Task Group hopes that this report will be a useful tool for Military Staffs, the Defence Research community, and technical experts within procurement agencies of NATO countries, by helping them to define and meet military requirements.

This updated report is the result of the contributions of all former RSG.10 and actual IST-011/RTG-001 members, which represent ten NATO countries (Belgium, Canada, France, Germany, the Netherlands, Portugal, Spain, Turkey, the United Kingdom, and the United States). A list of the National Points of Contact and of the authors of this report is given in the Appendix.

Because speech and language technologies are constantly improving and adapting to new requirements, it is the intention of the Task Group to continue to update this document as required in light of new evolutions. Therefore the Task Group appreciates any comments and feedback on this report.

Information System Technology Research Task Group 001

CHAIRMAN

Dr. Timothy Anderson
Air Force Research Laboratory
AFRL/HECA
2255 H Street
Wright Patterson AFB, OH 45433-7022
USA

SECRETARY

Mr. Carl Swail
Flight Research Laboratory
Building U-61
Montreal Road
Ottawa, Ontario K1A 0R6
CANADA

BELGIUM
Dr. Stéphane Pigeon
Koninklijke Militaire School
Leerstoel voor Telecommunicaties/
Royal Military Academy
Renaissancelaan 30
B-1000 Brussels

FRANCE
Dr. Edouard Geoffrois
CTA/GIP
16 bis avenue Prieur de la Côte d'Or
94114 Arcueil Cedex

GERMANY
Ms. Christine Bruckner
Bundessprachenamt – SM1
Horbeller Strasse 52
50354 Huerth

PORTUGAL
Prof. Carlos Teixeira
INESC-ID, Lisboa
Spoken Language Systems Lab
L2F – Laboratório de Sistemas de Língua Falada
R. Alves Redol, 9
1000-029 Lisboa

THE NETHERLANDS
Dr. David A. van Leeuwen
TNO Human Factors
P.O. Box 23
3769 ZG Soesterberg
Kampweg 5
3769 DE Soesterberg

TURKEY
Mr. Ozgur Devrim Orman
TUBITAK-UEKAE, National Research
Institute of Electronics & Cryptology
P.K. 74
41470 Gebze. Kocaeli

UNITED KINGDOM
Mr. Paul Collins
Room G007, A2 Building
DSTL Farnborough
Hampshire GU14 0LX

UNITED STATES
Mr. John J. Grieco
AFRL/IFEC
32 Brooks Rd.
Rome, NY 13441

Dr. Marc Zissman
Information Systems Technology Group
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9108

Use of Speech and Language Technology in Military Environments

(RTO-TR-IST-037)

Executive Summary

Communications, command and control, intelligence and training systems are making more and more use of speech and language technology components: i.e. speech coders, voice controlled C2 systems, speaker and language recognition, translation systems and automated training suites. Implementation of these technologies requires an understanding of what performance is possible with the products that are available today and those that will likely to be available in the next few years.

As speech and language technology become more available for integration into military systems, it is important that those involved in system design and program management be aware of the capabilities and the limitations of present speech systems. They must also be aware of the current state of research in order to be able to consider what will be possible in the future. This will be very important when considering future military systems upgrades.

The material contained within this publication include presentations of the current state of the art and the current research topics in selected speech and language technology areas: assessment techniques and standards, speech recognition, speaker and language identification, and translation.

La mise en œuvre des technologies de la parole et du langage dans les environnements militaires

(RTO-TR-IST-037)

Synthèse

Les communications, le commandement et contrôle, le renseignement et les systèmes d'entraînement font de plus en plus appel à des composants issus des technologies vocales et du traitement du langage naturel : il s'agit de codeurs vocaux, de systèmes C2 à commande vocale, de la reconnaissance du locuteur et du langage, de systèmes de traduction, ainsi que de programmes automatisés d'entraînement. La mise en oeuvre de ces technologies passe par la connaissance des performances des systèmes actuels, ainsi que des systèmes qui seront disponibles dans quelques années.

Etant donné l'intégration de plus en plus courante des technologies vocales et du traitement du langage naturel dans les systèmes militaires, il est important de sensibiliser tous ceux qui travaillent dans les domaines de la conception des systèmes et de la gestion des programmes aux capacités, ainsi qu'aux limitations des systèmes de traitement de la parole actuels. Ces personnes devraient également être informées de l'état actuel des travaux de recherche dans ces domaines, afin qu'ils puissent envisager les développements futurs. Cet aspect prendra beaucoup d'importance lors de la considération d'éventuelles améliorations à apporter à de futurs systèmes militaires.

Les textes contenus dans cette publication comprennent des communications sur l'état actuel des connaissances dans ce domaine, ainsi que sur des travaux de recherche en cours sur certaines technologies de la parole et du langage, à savoir : les techniques et les normes d'évaluation, la reconnaissance de la parole, l'identification linguistique, et la traduction.

Chapter 1 – INTRODUCTION

Speech can be considered as the most natural means of communication between humans. But the acoustical signals produced for this primary purpose also carry accompanying information about the speaker (gender, identification), his or her state of emotion, and the language spoken. It is therefore not surprising that speech technology embraces a wide range of applications in the civil and military world.

Before going into details regarding the potentials and military applications of speech and language technology, an attempt should be made at giving a rather simple notion of what this relatively new field of linguistic activities is all about. In a broad sense, it has to do with processing of language in its spoken and written forms through the help of the computer. Some basic, commercially available applications of speech and language technology include:

- Speech recognition – recognizing a human voice utterance either spoken in single (or isolated) words or as a continuous phrase (connected words); dictation systems are a well-known example;
- Speaker identification – identifying an individual speaker among a group of speakers;
- Language identification – determining the language in which an oral utterance is produced;
- Topic spotting – techniques used to monitor verbal utterances or written texts for thematic contents of particular interest; a more simple form is keyword spotting;
- Speech synthesis – the generation of (artificial) speech sounds by a computer-controlled machine;
- Machine translation – fully automatic translation of a text from one language into another by a computer.

Defence-oriented applications introduce challenges that are not always met by commercially available systems. For example, automatic speech recognition used in military scenarios must be robust to adverse conditions. This is because many military situations involve difficult acoustic and mechanical environments such as high and variable noise, vibration and g-force levels. The IST-011/TG-01 and its predecessor group, the RSG.10, have studied these military specific problems since their inception in 1978.

The primary goal of this updated report is to describe the military applications of speech and language processing and the corresponding technologies available. The degree of performance achieved by the various speech and language technologies is also described, and examples of successful integration in military systems are given. Military applications are itemized in six categories in this paper:

- Command and Control;
- Communications;
- Computers and Information Access;
- Intelligence;
- Training (inclusive of language training);
- Multinational Forces.

The available LRE (Language Research and Engineering) technologies include Spoken Language systems, Language processing, and Interaction between systems. The functional relationship between military applications and the corresponding technologies required using speech or text as medium or control signal is shown in Table 1.1. A reference to the corresponding chapter and section of this document is provided as well. The specific areas of military applications focused on the use of speech technology are described in Chapter 2; the available technologies are presented in Chapter 3.

Table 1.1: Relationship between Military Applications of Speech and Language Processing and Available Technologies. Numbering refers to text paragraphs.

Available Technologies Military Applications		Speech Processing 3.1	Speech Coding 3.1.1	Speech Enhancement 3.1.2	Speech Synthesis 3.1.3	Speech Recognition 3.1.4	Speaker Recognition 3.1.5	Language Identification 3.1.6	Language Processing 3.2	Topic spotting 3.2.1	Translation 3.2.2	Understanding 3.2.3	Interaction 3.3	Interactive dialogue 3.3.1	Multi-model communication 3.3.2	3-D Sound Display 3.3.3
Command and Control	2.1		•	•	•	•						•		•	•	•
Communications	2.2		•	•							•					•
Computers and Information Accesss	2.3				•	•					•	•		•		
Intelligence	2.4		•	•		•	•	•		•	•	•		•		•
Training	2.5				•	•								•	•	•
Joint Forces	2.6		•	•							•	•				
Case studies	5.0															
Cockpit Fast jet	5.1		•	•	•	•									•	•
Helicopter	5.2		•	•	•	•									•	•
Sonar	5.3					•										•
Noise reduction	5.4		•	•												
Training of air traffic controllers	5.5				•	•								•		•
Spoken Language Systems demonstration	5.6				•	•		•			•	•		•	•	
Battlefield Battle Management System	5.7				•	•				•				•	•	•
Speech Coders 600-1200 Bps	5.8		•	•												

For successful technical realization it is of great importance that the military requirements and specifications match with the performance offered by the systems. Adverse conditions prevailing in military activities, such as poor communications, noise environments, vibration, stress on operators, involvement of non-native speech, may deteriorate a speech signal or a text passage. Pertinent assessment and evaluation considerations are described in Chapter 4. Some practical applications are presented in Chapter 5.

Chapter 2 – SPEECH AND LANGUAGE IN MILITARY APPLICATIONS

The key military applications areas for speech and language technology, as indicated in Table 1.1, are: command and control; communications; computers and information access; intelligence; training; and multinational forces. In principle, all the speech and language technology areas are of some relevance to all fields of application, but Table 1.1 highlights particularly important relations between applications and technologies. The purpose of this chapter is to briefly discuss each of the fields of applications from the point of view of the requirements placed upon speech and language technologies.

In the multinational NATO context, combined or coalition military operations make the need for multilingual speech and language technology particularly important. For example, speech recognizers must operate in the languages of all the forces involved. In many situations it will also be necessary to process linguistic information (speech or text) in one language and to provide information, which is translated into another language or languages.

This need for multilingual operation adds to the usual military requirements for security, robustness against noise and jamming, and limited bandwidth channels. Moreover, speech communications systems must provide a high degree of performance despite the fact that many of the talkers and listeners will be working in languages, which are not their native language. This causes difficulties due to foreign accents and due to human comprehension limits in non-native languages.

The following sections outline military applications and associated requirements in each of the six fields listed above. In summary:

- Command and Control can be aided by human interaction with computers, weapons and sensor systems by voice. But this application requires high levels of performance of speech and language technology in real time, under adverse conditions including motion and noise and various effects of stress, and with multilingual input and output.
- Communications must operate securely, with high intelligibility, under conditions of noise and jamming. The speech signal, for example, must be coded and transmitted with sufficient fidelity to be understood by listeners who are not native in the language being spoken.
- Computers and Information Access are a crucial part of modern military operations. Speech and language technology can be used to allow military personnel to command and query computers and information by voice. This is particularly important for hands-busy, eyes-busy, and overloaded tasking of personnel. The diverse coalition environment also places new requirements on multilevel security for information systems. Access to information needs to be restricted to proper levels. Requirements on technology include speaker verification, audio data watermarking, multilingual input, and the possibility of translation or summarization of the information from one language into another.
- Intelligence collection places high demands on information processing and dissemination. In order for intelligence to be useful it must be information of high value, accurate, and reported in a timely manner. Due to an explosion of growth in the communications industry and the Internet as well as other open sources a steadily growing flood of audio and text data from multiple locations is available. This data must be filtered because of its' potentially high intelligence value. However an overload of available information has occurred which places high demands on filtering and information processing, including processing of speech and text.
- Training of forces for military operations can be significantly aided by applying speech technology to allow people to interact with advanced simulation systems by voice. In addition,

for multinational operations, training in foreign languages is essential; such training can be aided by utilizing speech and language technology to provide computer-assisted foreign language learning and tutoring for military personnel.

- Multinational Forces operations require the coordination of forces speaking different languages. Here, speech and language understanding and translation technology have great potential to enhance the efficiency and success of operations. However, the demands on technology are high, and initial applications probably need to focus on limited domains for translation and multilingual information exchange, with standardization of terminology and phraseology as a prerequisite.

2.1 COMMAND AND CONTROL

Command and Control activities concentrate on human operations. However, optimum functional behavior can be assisted by advanced interactive systems. A fully automated weapon system is neither feasible nor credible. One of the reasons for this is that it will never be as efficient as a human operator, in particular as far as adaptability to unknown situations, analysis of subtleties or details are concerned. Consequently the human operator will have to perform a major role in the basic command and control loop: perception – processing – decision – action. This loop is presented in Figure 2.1.

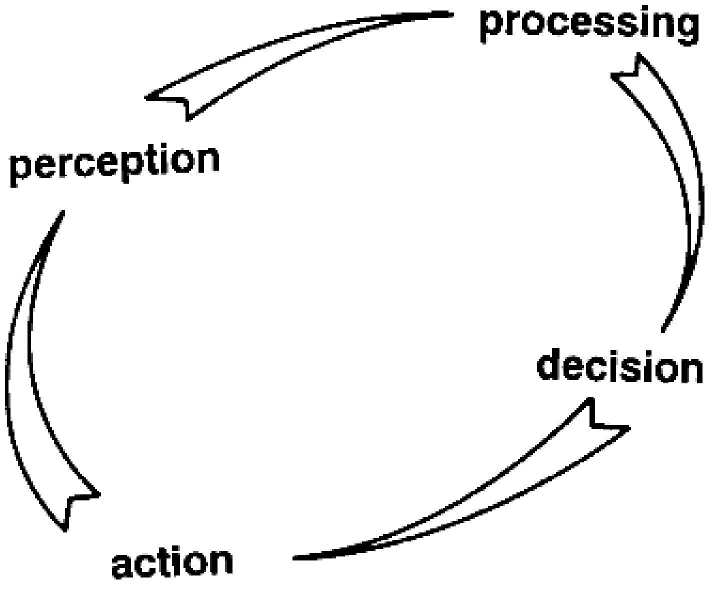


Figure 2.1: Control Loop of Human Decision Making.

On the other hand, the fact of having an operator in the loop always implies limitations with regard to the handling of a situation or a system. This can be defined by both the physical characteristics and abilities of a human operator and by his limited capacity to process information and to respond to it. For example, the body of the operator is sometimes crucial for the design of a technical system. If a tank or a cockpit is designed one has to take into account the fact that a person has to fit inside. For any system design one knows that a person has limited instruments of interaction (arms, fingers, legs, ears, eyes) sometimes impaired due to illness, stress or fatigue. Moreover, in order to be as efficient as possible, an operator has to be well trained, preferably in a very short period of time.

Although there is no generic solution to reduce the effect of these limitations, speech may help in some specific conditions: control of systems, reduction of workload, and acceleration of training. Analysis of a generic weapon system results in the definition of four top-level functional domains:

- Mobility,
- Communication,
- Survivability,
- Specific work functions of the system.

For each of these functional domains, speech technology has a potential field of application. The framework given in Table 2.1 considers two kinds of applications that are representative of the various situations: a C3I station and the crew in a tank or aircraft. These two examples are described in terms of the four top-level functional domains mentioned above.

Table 2.1: Functional Analysis of Two Command and Control Examples

	C3I station	Ground vehicles / Aircraft
Mobility	Mostly stationary	Principal functions: * mission planning and control * navigation * air and ground vehicle control
Communication	The core functions of C3I: * communication with own forces * communication with other C3I stations	Three areas of communication: * internal, between crew members * external, with other units or commands * control of on-board systems
Survivability	* potential corruption of information, implies secure data links * self-destruction, preventing potential use by enemy	* intrinsic mobility (see above) * communications security and protection against jamming
Specific work function	Potential functions (depending on the force level the station is installed): * action planning * overall situation presentation * intelligence collection	Potential functions (depending on the mission assigned): * fire power / destructive power * reconnaissance * firing order

Even if all kinds of weapon systems can benefit from the addition of tools based on speech technology, this does not imply that speech will be useful for each of the identified levels or functional domains.

For example, it does not make any operational sense to drive a tank entirely by voice. Also, tracking tasks executed by voice are inferior in performance with respect to tracking by hand. Speech is not suitable for this, as it does not have the ability for immediate feedback and fine-tuning of a parameter. However, speech can be very useful to control discrete and non-critical functions; e.g. radio channel selection, checklist control, data input and management, etc.

Each potential application has to be properly analyzed in order to know if and to what extent speech can help, given the potentialities of speech technology. A thorough human factors analysis is therefore required.

Integrating speech into a system is not trivial. It has to be taken into account at the very beginning of the design of the overall interaction between operator and system. The questions to be examined may include:

- Why use speech? What are the benefits of speech from the operator's point of view: improvement in accuracy, shortening of response time, workload reduction? In particular, the benefit of speech versus any other means of interaction has to be assessed.
- For what functions? Which interactions are to be mediated through speech, bearing in mind the impact on security, workload and usefulness (it is not necessary to build a function which is 100 % speech-based if it is seldom used)?
- For which speakers? This implies to know whether one or several persons communicate with the system. Consequently, speech recognition in either speaker-dependent or speaker-independent mode has to be chosen. It is also necessary to know the intellectual level of the users, since it has an influence on the choice of the means of interaction between the user and the system.
- Benefit versus cost: What will be the total cost of the system, and what will be the extra cost incurred by integrating speech-controlled functions into the system? This cost should include not only design and development costs but also operator training and maintenance costs.

First conclusions drawn from various studies performed in the armed forces indicate that speech can be used for about 50 % of the tasks in a C3I station, but only for 25 % in a tank or aircraft. In particular, speech techniques are not yet robust enough to be used on functions critical to safety, where even a very low error rate can have dramatic consequences (for example, if "eject" is understood instead of "reject" in an aircraft). Therefore, speech interaction can be envisaged for non-critical tasks only. In this case, a recognition performance of 95 % is in many situations satisfactory and better than human performance.

2.2 COMMUNICATIONS

The first speech transmission with electronic telecommunication means was realized in 1876, when Bell and Gray established the first telephone link on copper leads. Since that time the use of telephony has become ubiquitous in the preparation and conduct of military operations. A dialogue between two persons is more advanced and leaves less room for misinterpretation than the exchange of written messages. Presently wide band and narrow band communications are used. Written texts and complete files can be transferred by means of electronic data links. However, personal communications with the possibility of a dialogue or immediate response is of crucial importance in many operational situations. Speech communication is spread all over a military organization at all levels, but not all communications links are identical. The required level of security, the available radio link, the possibility of being jammed are all important factors, which determine system design. Two levels of communications can be identified, the staff level and the tactical level:

1) For inter-service (i.e. joint) and headquarters communications the following features are of primary interest:

   - The possibility to establish long distance communications;
   - Good intelligibility, even if the speech signal is transmitted through heterogeneous transmission media (guided waves on coaxial cable or on optical fibers, radio propagation above the ground, above the sea or in the air; free space propagation between earth stations and satellites);
   - The possibility to reach heterogeneous terminal platforms (fixed headquarters; mobile headquarters; ships; aircraft);
   - The use of conference and multimedia facilities;
   - Transmission of secure speech by signal encryption;
   - The use of translation facilities at both ends of the communication link.

All these features will lead to the choice of a network structure making use of several transmission means (telephone cable, optical fibers, line-of-sight links, satellite links). The speech signal has to be coded in a digital form to maintain a constant signal-to-noise ratio, independent of the length of the path. It is therefore mandatory to use the same coding technique at both ends of the communication link. This leads to the necessity of using common standards for speech coding. Such standardization is accomplished within NATO by standardization agreements (STANAGs). To reduce the bit rate and the required bandwidth, future systems (especially those on satellite links) will use coding techniques with "vocoders" that are based on the speech production parameters of the human vocal tract. These techniques could be combined with other services that also require an analysis of the speech signal. These new services are described in chapter 3. New trends in the integrated networks consist of merging digital speech signals with other data signals coming from different kinds of information sources (data, video, fax...). It is now possible to realize simultaneous communication of speech, graphics or image with multimedia terminals.

2) Speech communication for intra-service or tactical use.

In comparison with the users of Joint Staffs or the Main Headquarters a user of speech communications in a tactical unit is interested in the following features:

- The use of mobile light-weight terminals;
- The possibility to reach many units simultaneously (broadcast transmission);
- An acceptable quality of speech even in the presence of electronic countermeasures;
- The possibility to protect the spoken message by encryption.

These features lead to the use of radio nets that ensure the broadcast of messages. For military applications it is very important to send messages at the same time to different units. The use of the radio to transmit speech signals is very convenient for small units that have to move without constraints.

To protect these nets against electronic warfare (EW) threats, spread spectrum communication systems, like frequency-hopping nets, are used. These new techniques are not compatible with analogue signals. Accordingly, the speech signal has to be converted into a digital signal. To guarantee compatibility with inter-service (joint) communications, it is mandatory to use the same coding scheme, and eventually the same encryption method, in all speech terminals. Advances in jamming technology have produced systems, which are increasingly capable of disrupting und degrading critical communications. As a result, development and application of speech/audio technology for jam and tamper resistant communication systems is essential in overcoming improved jamming capabilities. This technology is critical for effective battlefield communications and overall success. Future developments of new speech compression techniques will provide ten to fifty times greater jamming resistance than current capabilities. Although the use of these techniques may place constraints on operations (e.g. limitation of the size of the vocabulary), these compression capabilities will provide an effective protection against jamming as well as improve operation in the presence of self-jamming. Effective utilization of our own EW resources requires Electronic Support Measures (ESM) designed to obtain the identification, location and disposition of opposing forces. Speech processing technologies such as speaker, language, and word recognition can augment other technologies and improve the identification and tracking of enemy forces.

2.3 COMPUTERS AND INFORMATION ACCESS

All levels of military operations now require human interaction with computers and with databases for entry or retrieval of information. Applications ranging from command and control, to logistics and maintenance, and to forward observer reporting, all will require computer access. Speech technology has great potential in these applications for allowing military personnel to query or command computers by voice.

Multilevel access to classified material is also a concern. Along with the ability to store vast amounts of information in automated information systems comes the responsibility to protect the information. Individuals without the proper security level need to be restricted from certain data. This requires the need for an access method using biometrics. Voice biometrics can be particularly efficient in this field. Speaker verification technology has great potential for accomplishing this task.

The decreasing cost and increasing power of computer hardware and software have made the use of computers universal in the armed forces. In addition, the quantity and range of types of available digital information have increased dramatically, to include not only text and numerical data, but also sound, graphics, images, and video. It is now possible to access and manage this kind of information with small, low-cost terminals. For example, the existence of digital terrain maps and compact presentation systems allow military aircraft pilots and navigators to make direct use of this type of information during their missions.

Now that we have more computers, more information, and more capability to access and process this information, the man-machine interface can become a bottleneck. Typically, people control access to the computer by keyboard, mouse, trackball, and touch panels. Output information is typically presented by displays or printers. But this type of input/output requires the use of the hands and eyes, which in military operations are often busy with other tasks. For example, it would be difficult for a forward observer to watch the target and at the same time enter data via keyboard.

For such military operations, speech recognition can provide a useful input mechanism, while speech synthesis, i.e. computer-generated speech, can provide a useful output mechanism.

Requirements on the technology for these applications include: multilingual speech recognition and synthesis, and high performance recognition under conditions of stress, workload, and noise. Specific applications include: repair and maintenance; control of computerized auxiliary systems; report entry for forward observers; and access to logistics databases.

Besides providing help in man-machine interface problems, language technology can be very useful for converting information into a form, which is understandable to the user. For example, machine translation technology provides the possibility of translating or summarizing information, which is stored in a language foreign to the user, into the user's native language. Such multilingual information processing represents both a challenge and an opportunity for advanced language technology.

2.4 INTELLIGENCE

Intelligence requires processing of a large variety of types of information, including speech and text in numerous languages. With the growing complexity of a multi-polar world, the number of languages in which information needs to be processed continues to increase. Also, the rapidly changing nature of the world situation requires that information processing be able to adapt quickly to new languages and new domains.

Experience gained from crisis reaction, peace support and peace enforcement operations, in particular the lessons learned in the Former Yugoslavia and the Kosovo, has underlined the extraordinary importance of efficient reconnaissance and intelligence activities. Moreover, the Information Age with the advent of the new media and its overwhelming flood of information have created a mass data problem: the vast amount of information available from open sources (press, broadcast, television, Internet, etc.) requires an enormous effort of analysis work which meets rather soon with its limits when it comes to the time and manpower factor. Mitigation of the problem with acceptable levels of reliability might be achieved through the use of increasingly automated speech and language technology tools and procedures which now are becoming available (keyword spotting, topic spotting, contents extraction, summarization, translation, etc.).

As shown in Table 1.1, essentially all of the available speech and language technologies have potential applications to this information processing activity.

2.5 TRAINING

Well-trained personnel are imperative for the success of military missions. The definition of requirements for language and speech should take into account the following aspects:

2.5.1 Language Training

It must be differentiated between active and passive use of spoken and written language, e.g. in C3I activities. New applications of speech and language processing are able to support language training, using:

- Computer-assisted learning systems;
- Production of computer-aided test facilities;
- Production of didactic material.

2.5.2 Simulation

The need for using simulators in weapon systems training stems from the immense costs involved in live exercises. Therefore, speech and language applications must be integrated at all levels of simulator-based training. In particular, verbal man-machine dialogue, man-machine interfaces, voice-activated system control and feedback will play a growing part in training activities.

2.5.3 Requirements

The following requirements emanate from the above-mentioned facts:

- Reduction of training time and cost;
- Increase of number of languages to be handled;
- Maintaining and improvement of proficiency levels;
- Tele-learning.

2.6 MULTINATIONAL FORCES

In 1994, at the Summit Meeting of the North Atlantic Council held in Brussels, decisions were taken which led to the development of the Combined Joint Task Forces (CJTF) concept. The need for such a concept – enabling the participation of non-NATO States in crisis operations (thus making the language problem even more complex) – arose from the changing security situation in Europe and the emergence of smaller but diverse and unpredictable risks to peace and stability. In particular, it was agreed that future security arrangements would call for easily deployable, multinational, multi-service military formations tailored to specific kinds of military tasks. These included humanitarian relief, peacekeeping and peace enforcement missions, as well as collective defence. The forces required would vary according to the circumstances and would need to be generated rapidly and at short notice.

The integration of military units coming from a number of countries into Combined Joint Task Forces or similar multinational formations can cause a lot of language understanding problems at different stages of the coalition-building process:

- During the negotiation of the integration plans;
- During the preparation of training exercises;
- During the generation of mission orders;
- During the performance of the mission.

These problems arise due to the lack of:

- Basic knowledge of the foreign languages involved;
- Common definitions of critical military terms (standardization of terminology);
- Employment of working automatic machine translation systems.

Therefore, the players in the multinational joint forces theater of operations need to be assisted by new means mainly based on speech and language technology, e.g.:

- Automatic translation of mission orders and messages to considerably reduce the reaction time of the military units;
- Conversion of voice reconnaissance data transmitted by a human observer into language-independent data;
- Automatic generation of specific military glossaries and multilingual dictionaries to be used at all staff and operational levels.

The major benefits of introducing speech and language technologies in a multinational military environment will be:

- Enhancing the mutual understanding process;
- Increasing the speed of operational exchange of multilingual information;
- Reducing the reaction time of operational units in critical situations;
- Speeding-up of foreign language acquisition and learning.

Chapter 3 – AVAILABLE TECHNOLOGIES

Speech technology is traditionally divided into speech processing and natural (i.e. written) language processing. This chapter presents these and their combination, followed by some related technologies.

3.1 SPEECH PROCESSING

Modern speech technology is based on digital signal processing, probabilistic theory and search algorithms. These techniques make it possible to perform significant data reduction for coding and transmission of speech signals, speech synthesis and automatic recognition of speech, speaker or language. In this section the state-of-the-art is presented and related to realistic military applications.

3.1.1 Speech Coding

When digital systems became available, it was obvious that the transmission of digital signals was more efficient than the transmission of analogue signals. If analogue signals are transmitted under adverse conditions, it is not easy to reconstruct the received signal, because the possible signal values are not known in advance. For digital signals discrete levels are used. This allows, within certain limits, the reconstruction of distorted signals. The first digital transmission systems were based on coding the waveform of the speech signal. This results in bit rates between 8000 to 64000 Bps (bits per second). The higher the bit rate the better the quality. Later, more advanced coding systems were used where basic properties of the speech were determined and encoded, resulting in a more efficient coding (bit rates between 300 and 4800 Bps) but also in reduced intelligibility. These methods are discussed in this section.

The first technique used to convert an analogue signal into a digital signal was based on the work of Shannon. He converted the instantaneous signal value at discrete moments into a binary number (a sample) and proved that it was possible to reconstruct the original signal from these samples, if the sampling frequency was high enough. Theoretically the sampling frequency is required to be twice the highest frequency component of the analogue signal.

Based on this technique a conversion system was used for telephone speech signals (with a frequency range between 300 and 3400 Hz) by using a sampling frequency of 8 kHz. The conversion of the instantaneous signal value had a resolution of 256 discrete levels corresponding to 8 binary digits. These bits were then transmitted in series with a bit rate of 8 × 8000 Bps or 64000 Bps. This technique is known as Pulse Code Modulation (PCM), and is still in use. PCM is one of the methods used to realize time division multiplex (TDM) where bit streams of different channels are combined in order to transmit many simultaneous telephone links using the same transmission channel.

In order to reduce the bit rate, and thereby increase the number of simultaneous channels in a given bandwidth, it was necessary to increase the coding efficiency. Therefore, the signal was compressed before encoding at the transmission end, and expanded after decoding at the receiving end. There are presently two different compression algorithms in common use: the so-called A-law used in Europe, and the μ-law used in North America. The differences between the two methods are small and it is possible without much distortion to use one of the two methods at one end and the other method at the other end.

Another method to convert analogue speech signals consists of using a delta-modulator. In this case, the sampling frequency is much higher than twice the highest frequency component and only one bit of information is transmitted per sample, corresponding to the slope of the signal (differential quotient). By making use of a simple integrator, the original waveform can be retrieved. This technique results in good signal quality at lower bit rates than is required for PCM. In general a bit rate of 16 or 32 kBps is used.

Further enhancements to these methods, including dynamic optimization, have resulted in the CVSD (Continuous Variable Slope Delta modulation) and ADPCM (Adaptive Differential Pulse Code Modulation) methods.

The relation between the instantaneous analogue value of the waveform and the digital representation is different for PCM and Delta Modulation. For PCM, the most significant bit of the digital representation represents the biggest portion of the analogue value; hence digital errors are more dramatic if this value is distorted. For Delta Modulation, all bits have an equal significance, making this method more robust to channel errors. PCM error rates of 1% will give an unacceptable deterioration, while for Delta Modulation error rates up to 15% will result in an acceptable quality.

Whereas waveform coders like the ones described above aim at a faithful reproduction of the signal waveform, vocoders explore our knowledge of speech production, attempting to represent the signal spectral envelope in terms of a small number of slowly varying parameters. Vocoders achieve considerable bit rate savings, being aimed at bit rates below 2400 Bps; however, they result in degradation of the speech quality and of the ability to identify the speaker.

Many new coding schemes have been proposed recently which cannot be classified according to the waveform-coder/vocoder distinction. This new generation of coders overcomes the limitations of the dual-source excitation model typically adopted by vocoders. Complex prediction techniques are adopted, the masking properties of the human ear are exploited, and it becomes technologically feasible to quantize parameters in blocks (VQ-vector quantization), instead of individually, and use computationally complex analysis-by-synthesis procedures. CELP (Code Excited Linear Prediction), multi-pulse, and regular-pulse excitation methods are some of the most well known “new generation” coders in the time-domain, whereas those in the frequency-domain include sinusoidal/harmonic and multi-band excited coders. Variants of these coders have been standardized for transmission at bit rates ranging from 12 down to 4.8 kBps, and special standards have also been derived for low-delay applications (LD-CELP).

Today, audio quality that can be achieved with the so-called telephone bandwidth (3.4 kHz) is no longer satisfactory for a wide range of new applications demanding wide-band speech or audio coding. At these bandwidths (5 – 20 kHz), waveform coding techniques such as sub-band coding and adaptive transform coding, have been traditionally adopted for high bit rate transmission. The need for 8-to-64 kBps coding is pushing the use of techniques such as linear prediction for these higher bandwidths, despite the fact that they are typical developed for telephone speech. The demand for lower bit rates for telephone bandwidth is, however, far from exhausted. New directions are being pursued to cope with the needs of the rapidly evolving digital telecommunication networks. Promising results have been obtained with approaches based, for instance, on articulatory representations, segmental time-frequency models, sophisticated auditory processing, models of the uncertainty in the estimation of speech parameters, etc. The current efforts to integrate source and channel coding are also worthy of mention.

Although the main use of speech coding so far has been transmission, speech encoding procedures based on Huffman coding of prediction residuals have lately become quite popular for the storage of large speech corpora. For an exhaustive survey of speech coding, see Spanias (1994), which includes over 300 references to the relevant literature and comparisons of different schemes from 64 kBps down to 800 Bps, on the basis of MOS and DRT scores (see section 4.2) and computational complexity.

Two concerns, security enhancement and bandwidth reduction drive the applications of speech coders in military applications. The bandwidth considerations (and to some extent security) are driving similar interests in the commercial area, in particular in mobile telephones. Wireless communication is further driving the development of speech coding algorithms. The same needs are found both in tactical military and new commercial systems. The requirement to go from a hand-held terminal directly to a satellite imposes power and antenna size considerations that demand low bit-rate communication. Bit rates in the

2400 to 4800 Bps are typical of these requirements. Since these systems will often be used in mobile or other noisy environments, the algorithms must be robust to acoustic background noise and operate over fading communications channels.

3.1.1.1 Current NATO Standards

NATO has three Standardization Agreements (STANAG) in the speech coding area. The first is STANAG 4198, titled "Parameters and coding characteristics that must be common to assure interoperability of 2400 bps linear predictive encoded digital speech". STANAG 4198 defines speech coding characteristics, the coding tables and the bit format requirements to ensure compatibility of digital voice produced by the linear predictive coding (LPC) technique at 2400 bps rate [NATO STANAG 4198]. In their basic form, Linear Predictive Coding (LPC) algorithms for speech achieve high compression ratios by developing short-term, steady-state models of the vocal tract and transmitting only the quantized and encoded parameters of these models. The speech production process is modeled by a flat spectrum excitation source representing the glottal movement, which is filtered by an all-pole short-term stationary digital filter representing the shaping due to the response characteristics of the vocal tract. This STANAG has been implemented in many strategic and tactical secure voice systems such as Narrow Band Secure Voice System (NBSVS), Tactical NBSVS, Secure Terminal Unit 2 (STU-2), Advanced Narrow Band Digital Voice Terminal (ANDVT) and HF (High Frequency) radios by NATO nations.

STANAG 4479, titled "Parameters and coding characteristics that must be common to assure interoperability of 800 bps digital speech encoder/decoder", was prepared for frequency hopping HF radios [NATO STANAG 4479]. The core of STANAG 4479 is based on LPC 10e speech coding methodology at 2400 bps. By using vector quantization, the bit rate of the vocoder was reduced to 800 bps.

STANAG 4209, titled "The NATO Multi-Channel Tactical Digital Gateways Standards for Analogue to Digital Conversion of Speech Signals", also defines the Continuously Variable Slope Delta-modulation (CVSD) speech coding algorithm at 16 kbps. The CVSD algorithm attempts to reconstruct the exact waveform at the receiver that was input to the transmitter; thus it is classified as a waveform coder. The digital output from the transmitter is 1 bit per input sample up-sampled to 16Kbps. The transmitted bit stream is used to indicate the slope of the input waveform. The slope-limiting detection logic looks at the 3 most recent bits transmitted. If these bits are all 1's or all 0's, the step size is doubled. For all other combinations, the step size is cut in half. The ratio between maximum and minimum step size is 16. The sign of the slope is positive if the current bit is a 1 and negative if the current bit is 0. The CVSD algorithm has been implemented in UHF/VHF secure voice equipment such as KY-58/57, SATCOM voice terminals and digital gateways.

NATO legacy strategic/tactical secure voice communications systems use either the 2.4Kbps Linear Predictive Coding (LPC10e), or the 16Kbps Continuously Variable Slope Delta Modulation (CVSD) algorithm for speech compression. These algorithms were considered the state of the art in narrow band speech coding when introduced 20 to 30 years ago.

3.1.1.2 New NATO Standard on Speech Coding

NATO forces rely on secure voice equipment provided by the member nations to satisfy joint interoperable secure communications requirements. NATO military communications require high quality, reliable, interoperable communications that support both the tactical and strategic commanders as they pursue NATO military missions. Seamless interoperability and end-to-end security between the domains of strategic, tactical, SATCOMs, and internetworking protocols is of paramount importance. An important element of their overall communications requirements is voice communications. Currently, the performance of in-place voice coding algorithms is unacceptable in the harsh tactical acoustic conditions where NATO Commanders operate (such as tracked vehicles and helicopters). These legacy speech coding algorithms now represent obsolete technology and require tandem operations, which severely degrade the Quality of Service (QoS), provided to the users [Test and Selection Plan, 2000].

A new generation of 1.2Kbps/2.4Kbps speech coding algorithms has been developed which far exceed the QoS of the existing speech coding algorithms. Furthermore, these new speech coding algorithms provide superior performance in harsh acoustic noise conditions, improved speaker recognizability, and improved non-native speaker intelligibility. Additionally, the new speech coding algorithmic suite represents the first enabling technology for seamless interoperability and end-to-end security across the user and network domain boundaries of the strategic and tactical environment as firmly expressed in MC-337.

NATO is currently undergoing a competition among the nations to select the next generation secure voice coding STANAG 11 for the future. The Ad-Hoc Working Group on Narrow Band Voice Coding (AHWG-NBVC) is tasked with the selection of a new voice coding algorithm STANAG at 1.2Kbps and 2.4Kbps. Future NATO secure voice equipment requirements should include these algorithms as a minimum essential interoperable mode of secure voice communications. These speech coding algorithms will enable broad interoperability across the entire NATO communications infrastructure while providing superior performance.

The new STANAG contains a primary speech coding algorithm at 2400 bps and a secondary speech coding algorithm at 1200 bps. The new speech coding algorithms should provide superior performance and/or improve an existing Quality of Service (QoS) in strategic and tactical voice communications in terms of speech quality, speech intelligibility, robustness to ambient noise and channel errors, communicability, tandem sensitivity, speaker recognizability, and language dependency [Decision Sheet-2]. Both rates must be based upon the same core technology, but differ only in their quantization processes necessary for transmission. The maximum delay is not to exceed 250 ms for either rate. [Decision Sheet-5] Additionally, both speech coding algorithms must be compatible with existing and future planned crypto systems by providing a sync bit in every frame [Program of Work].

The new STANAG speech coding algorithms shall be freely available for use by MNCs and NATO member nation military forces (including Partnership for Peace nations) and for use by military coalition forces (including non-NATO national forces requiring interoperability) sanctioned by NATO and/or the United Nations in future peacekeeping missions. Other use by any nation may be subject to IPR restrictions and/or licensing fee arrangements [Program of Work].

The new STANAG is assigned STANAG 4591 and titled "Parameters and coding characteristics for interoperability of NATO Post-2000 digital voice coding at 2.4 Kbps and 1.2 Kbps" [Decision Sheet-9].

3.1.2 Speech Enhancement

A technology that is of high interest to both military and civilian voice communication developers and users is speech enhancement. We define enhancement here as any process that improves the quality and/or the intelligibility of the speech signal for human listening and/or machine processing. This includes not only pre and post emphasis enhancement techniques, but also the reduction or the removal of any co-existing signal other than the desired speech signal to include both distortion and noise.

The interest in speech enhancement continues to be high because of the increased need to remove man-made contamination. Although improved electronic equipment continues to reduce the noise generated by electronic devices, steady increases in the density of signals in the radio frequency and acoustic environment, along with increases in jamming techniques and equipment, have dramatically increased information and non-information bearing man-made contamination. Because there are many sources of contamination that can reduce the quality and intelligibility of speech, and because there continues to be a steady increase in man-made contamination, speech enhancement technology has been and continues to be pursued in both development and application.

As a result, many types of enhancement processes have been and continue to be developed to address a wide variety of contamination types encountered in every day voice communications. To be able to effectively apply enhancement technology, it is necessary to identify the source and category of contamination being encountered and then select and apply the appropriate enhancement process.

To aid in this process, we have partitioned the sources of contamination into four categories: Electronic Device, Natural, Non-Information Bearing Man Made, and Information Bearing Man Made. Table 3.1 shows some example sources for each contamination category.

Table 3.1: Contamination

Contaminating Source	Categories of Contamination			
	Electronic Device	**Natural**	**Non-Information Bearing Man Made**	**Information Bearing Man Made**
	- Thermal Noise - Shot Noise - Inter-modulation Distortion	- Galactic Noise - Atmospheric Noise - RF/AF Propagation	- Ignition Noise - Corona Noise - Acoustic Noise - RF/AF Jamming	- Radar - Radio - TV

Each of these sources of contamination produce its own type of noise or distortion such as static, hums, unwanted speech, clipping, etc. Since each type has its own characteristics, no one speech enhancement process can cope with any one contamination category or noise source. There are just too many contaminating types and characteristics. Hence, there have been a large variety of enhancement processes and techniques developed to address each source or type of contamination. A discussion of enhancement techniques follows.

3.1.2.1 Enhancement Techniques

Speech enhancement techniques can be partitioned into four classes. They are:

1) Acoustic Isolation and Cancellation where the speech signal is acoustically enhanced by shielding, absorbing and/or canceling an undesired acoustic signal.
2) Signal Level Control used to enhance the speech signal by controlling the amplitude to minimize distortion and improve SNR.
3) Electronic Noise Reduction and Cancellation which enhances the speech signal using analog electronic equipment and/or signal processing filtering and cancellation techniques.
4) Speech Manipulation to enhance the speech signal by time and/or frequency modification.

Each of these enhancement classes has multiple techniques that can be applied to a given application. Table 3.2 shows the types of enhancement techniques used for each class. A discussion of these classes and techniques is provided here to familiarise military users with the technology and provide insight on where to apply it.

Table 3.2: Enhancement Techniques

Acoustic Isolation Cancellation	Signal Level Control	Electronic Noise Reduction	Speech Manipulation
- Handphones - Headsets - Shielding Materials - Microphones - Mic Arrays - Windscreens - Absorption Materials	- Signal/Voice Limiting - Speech Compression - AGC/ALC	- Impulse - Tonal - Wideband - Interfering Speech	- Spectral Weighting (pre-emphasis, de-emphasis, channel normalization) - Speech Time Scale Modification (rate changes, pitch normalization)

3.1.2.1.1 Acoustic Isolation and Cancellation

These enhancement methods address the reduction or elimination of environmental acoustic contamination. Examples of military environment with severe acoustic contamination include the aircraft cockpit and inside a tank. To reduce these acoustic noises for voice communications, enhancement techniques, such as acoustic shielding, absorption, and cancellations are applied. The most common shielding method is the use of headphones. The isolation provided by wearing headphones is very effective at attenuating frequencies above 200 Hz.

Large environmental noise below 200 Hz, such as in a cockpit, requires active noise reduction (ANR) headsets. ANR headsets use electronic noise cancelling circuitry to reduce environmental noise. Figure 3.1 shows how effective this technology can be in reducing environment acoustic noise.

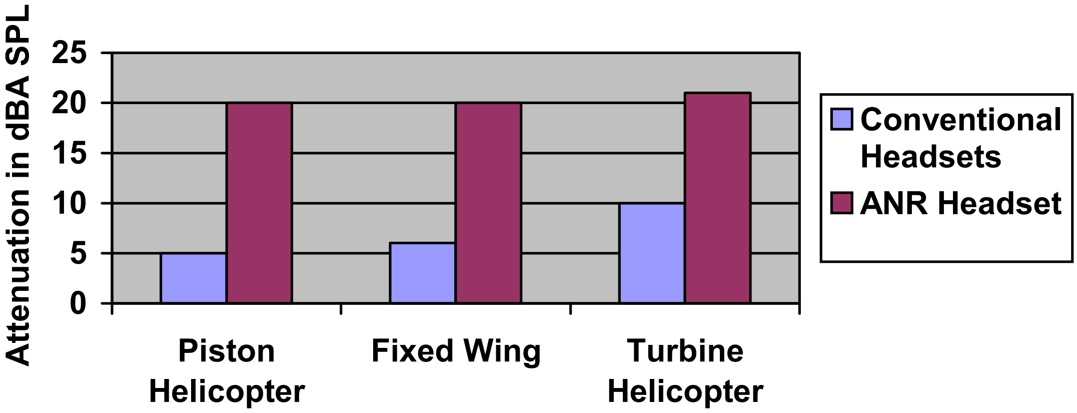


Figure 3.1: Attenuation Obtained by Conventional and ANR Headsets.

Other speech enhancement techniques used to reduce acoustic environmental noise are accomplished at the microphone. These techniques, which include noise cancellation microphones, microphone arrays and wind screens, are used where strong environmental noise contaminates the speech signal. Typical attenuation that can be obtained by these techniques varies with frequency and can range from approximately 5 to 25 dB. However, reduction in noise level does not always result in equivalent improvement in intelligibility. In Figure 3.2, the transmission quality, expressed in STI (Speech Transmission Index, an objective measure for intelligibility) is given for two types of microphones as a function of the environmental noise level. Good communication is obtained for an STI of .60 and above. The Figure shows that good communications can be obtained in a noise environment approximately 90 dB(A) for microphone B and approximately 110 dB(A) for microphone A, both at a 5 cm distance from the mouth. Performance drops off as the distance is increased between the microphone and the mouth, demonstrating the importance of microphone position for communications in high noise environments.

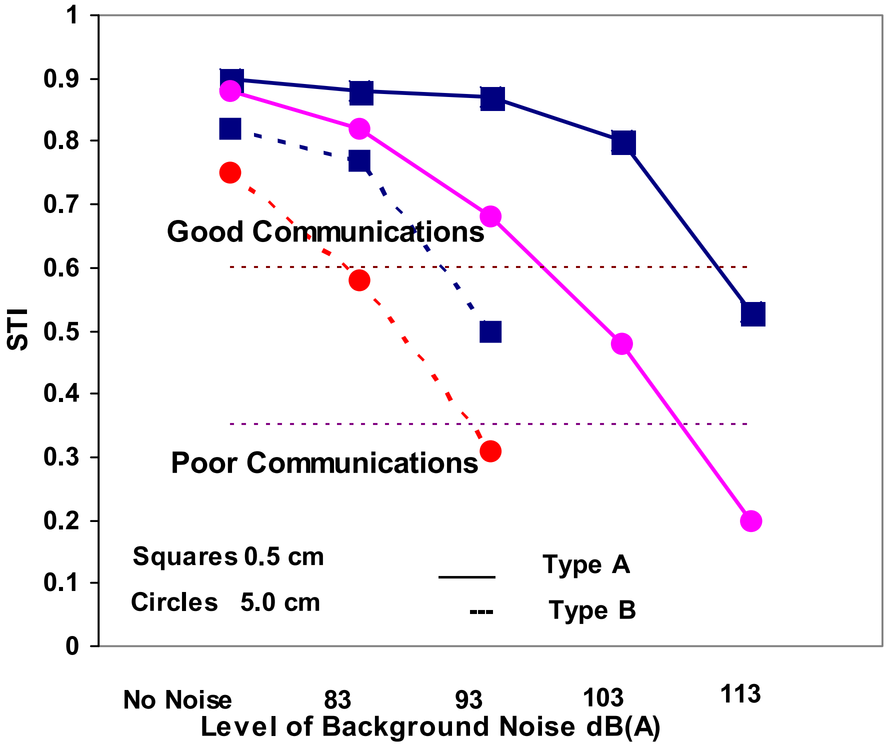


Figure 3.2: STI as a Function of the Noise Level for Two Different Microphones and Two Speaking Distances.

3.1.2.1.2 Signal Level Control (SLC)

Level control techniques are used to enhance speech by controlling the amplitude of the signal. SLC reduces signal distortion and attempts to increase signal-to-noise ratio. It is sometimes given little attention, but is a very important factor because it impacts on signal quantity and intelligibility in any voice communications process. There are three basic types of SLC – Signal/Voice Limiting, Speech Compression, and Automatic Gain Control. All techniques attempt to prevent hard clipping (the signal is flat topped on peaks) of the signal which in turn reduces the harmonic distortion that clipping causes.

Signal /voice limiting uses very non-linear amplitude response on the strongest or highest energy parts of the signal. Figure 3.3 shows a typical limiter characteristic. This technique is used where there is plenty of signal level and to reduce distortion of high level signals over heavy clipping of the signal. Limiting does not affect the low and medium signal levels.

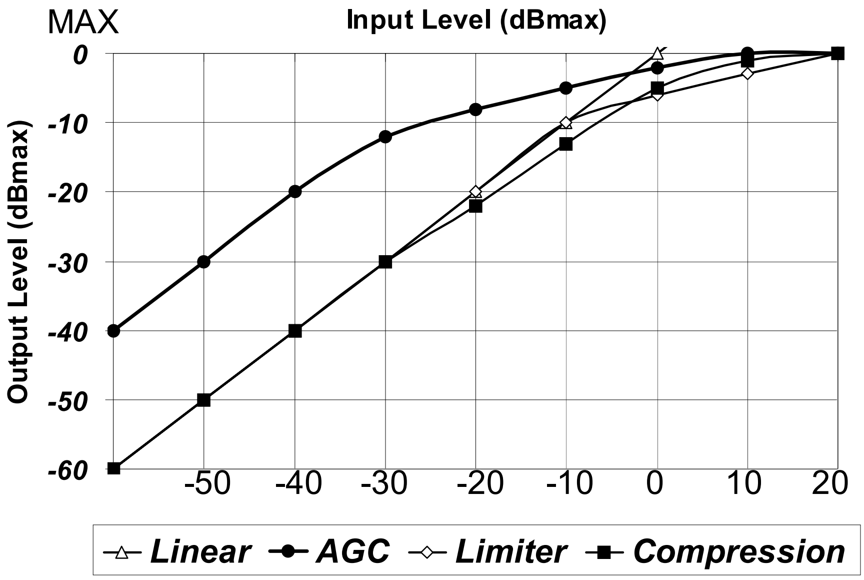


Figure 3.3: General Characteristics of Three Types of Automatic Level Control.

Another type of level control is speech compression. Speech compression techniques can increase signal-to-noise ratio and will produce a signal with less distortion to high level signals than signal limiting. The signal can also be expanded before listening or processing to remove the compression and its distortion effects.

A third type of level control used to enhance signals is Automatic Gain Control (AGC) sometimes called Automatic Level Control (ALC). It impacts on the low energy parts of the signal as opposed to the high-energy parts as does compression and limiting. It is used where the low energy parts of the signal need amplification with minimum distortion. Figure 3.3 shows the general characteristics of AGC.

Signal level control, if applied correctly, can enhance the speech signal for both listeners and machine recognizers. The type of SLC used is application dependent and can be selected with careful consideration of the type of information discussed above and shown in the figure.

3.1.2.1.3 Electronic Noise Reduction and Cancellation

Electronic noise reduction can be accomplished by using analog and/or digital signal processing techniques. The noises that are addressed may be stationary or non-stationary and have many shapes or colors. As a result, researchers have developed many types of noise reduction algorithms to address different noises. Both the noise and reduction methods can be partitioned into four classes – Impulse Removal, Tone Reduction, Wideband Reduction, and Interfering Speech Suppression. Each class of electronic enhancement is shown in Table 3.3 below with examples of the noises they reduce or remove.

Table 3.3: Classes of Electronic Enhancement

Electronic Enhancement Process	Examples of Noises Addressed
Impulse Removal	Ignition/engine noise, static, microphone clicks, corona discharge
Tone Reduction	Power line hum, power converter noise, heterodyne beats, data, on-off and FSK keyed signals
Wideband Noise Reduction	Atmospheric noise, electronic shot noise, record media noise
Interferring Speech Suppression	Adjacent, channel and co-radio interference, crosstalk, echo, reverberation

Impulse Noise Reduction

Impulse noise is a non-continuous signal consisting of short duration that may be periodic or random. Periodic impulses, such as that generated by an automobile's ignition or electric motor can be very effectively removed. Unlike these sources, impulses that are approximately the same amplitude as speech peaks and have duty cycles greater than approximately 30% are difficult to remove and impact on both speech quality and intelligibility. Further, no matter how good the detection and removal processes are, if quality and intelligibility are to be maintained or improved, it is essential that the impulse process replace the removed impulse area with a speech estimate that leaves no boundary discontinuities.

There are several methods that are used to fill the removed impulse area. All of them make estimates that are based on speech segments on each side of the impulse area. The processes that produce the best estimates give superior listening results. Figure 3.4 shows the response of a time domain process in removing large periodic impulses (>4 times the average speech level) as a function of impulse rate and duty cycle. Note that an impulse rate of 200 per second, which relates to an 8 cylinder automobile engine turning 3000 rpm is removed.

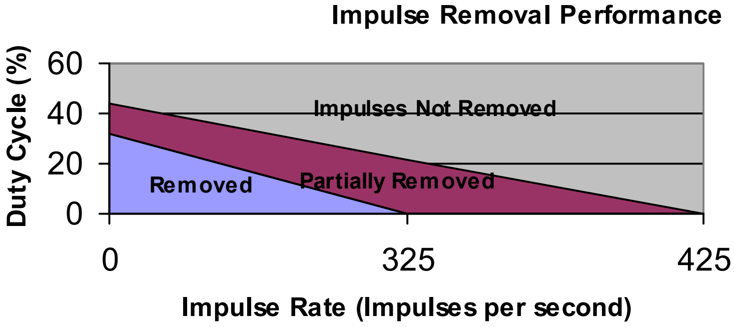


Figure 3.4: Impulse Removal as a Function of Impulse Rate and Duty Cycle for a 1 millisec Impulse Width.

Tone Reduction Technology

There are many techniques for reducing tonal contamination, such as power line hum, adjacent channel interference, tonal data signals, tonal jammers and others. Many types of tone reduction processes are required because of the large variety of tonal characteristics contained in these contaminating signals. For example, power line hum requires a process (usually comb filtering of some type) that attenuates many stationary harmonics of the power line frequency, while multitone data signals are not harmonically related and are keyed "on" and "off", making a comb filter approach ineffective.

Large amplitude tones are most easily detected by peak picking threshold methods and attenuated in the frequency domain. Small tones require a frequency integration process of some type and are again attenuated in the frequency domain. The limit for reliable detection of narrowband tones is dependent on how stationary the tones are in frequency and amplitude. When the interfering tones move at the rate of change of the speech of the speaker, the tone detection processes become unreliable since the interference can no longer be distinguished from the speech. If the moving tones are large in amplitude (several times larger than the speech) they can be reduced using peak picking threshold techniques at a rate of change of at least 250 Hz/second with little effect on the speech. Integration methods, such as a histogram process, are ineffective for moving tonal interference.

Wideband Noise Reduction

Wideband noise is a common contaminate of speech communications. It is often introduced as a contaminant by radio frequency and acoustic propagation, record processes, digital processes and electronic equipment. Wideband noise is generally random and has many colours with white and pink being the most common. It can be characterised as a hiss or rumble like sound to a listener. Because of its wide bandwidth, it overlaps the speech signal in both the time and frequency domain, making it a challenge to reduce without some impact on speech quality and intelligibility. Wideband noise reduction technologies attempt to reduce the noise by sampling the noise to obtain some type of noise estimate or model that is in some process subtracted from the noise plus speech signal. Generally increasing the amount of processing to obtain higher signal-to-noise ratio (SNR), increases the distortion imposed on the speech signal. These reduction processes can improve SNR anywhere from 10 to 25 dB, depending on the type of wideband noise, the original SNR, and the amount of distortion that is tolerable to the listener or machine recogniser. For intermittent wideband noise, such as in radio communications, it is important to consider the adaptation time because it can span anywhere from tens of milliseconds to seconds, depending on the type of reduction process and parameters used. Real-time reduction processes with long adaptation times will perform poor on short communication transmissions.

Reduction of Interfering Speech

There are many sources of undesired speech contamination and many interfering speech reduction processes. Interference from two or more speakers talking simultaneously (voice on voice) is commonly encountered and is one of the most difficult interference to reduce. In radio frequency communications, voice on voice is referred to as co-channel interference caused by two or more emitters on the same frequency.

In telephone communications and in audio recording, the same voice on voice interference is called crosstalk or channel bleedover. While humans can partially compensate for this interference, the performance of automatic speech processing systems, such as speech and speaker recognizers, deteriorates drastically in the presence of such interference.

Most techniques attempt to separate the speech of two talkers within each analysis frame using pitch-based separation methods. Pitch estimates are made and then used by a separation process (such as a comb-filter) to enhance the desired speaker's voice and/or to suppress the interfering speaker's voice. These methods have had limited success (10 to 15% improvement) under conditions of good SNR (>30dB).

Other types of speech interference involve only one speaker. Echo, reverberation and multipath are sources of interference where the talker's speech interferes with its self. All of the interferences are characterized as delayed and attenuated versions of the same speech. Multipath propagation is the radio communication term and echo and reverberation are acoustic terms, although echo cancellers are used and very effective in telephone communications.

3.1.2.1.4 Speech Manipulation

There are a variety of manipulation methods used to enhance speech. Manipulation in the frequency domain is called spectral weighting and is used to shape the spectrum to improve SNR and in turn human and machine recognition of words and phrases. Spectral weighting includes pre-emphasis, de-emphasis, and channel normalization. When these techniques are applied properly, improvements in recognition can be equivalent to a ~10dB improvement to SNR.

Manipulation of speech in time is called time-scale modification. Time-scale modification is an enhancement technique used to slow down the speaking rate to improve recognition and comprehension. Slowing the speech rate of a low intelligibility message can enhance critical acoustic cues and improve message understanding when pitch normalisation (a process for re-establishing the original pitch after slow down) is accomplished. State-of-the-art systems can change the rate of speech to as slow as .3 of the original speed. However, it has been shown that rates slower than .6 (of the original speed) provide little to no advantage. Near real-time implementations are possible for most audio manipulation techniques. These techniques can be used in tandem with other enhancement techniques to improve both human and machine recognition.

3.1.2.1.5 Enhancement Summary

There are many types of enhancement technologies and processes for a large variety of contamination types. To select a technology, it is necessary to identify the source and type of contamination being encountered. Referring to Table 3.2, along with the enhancement technology descriptions, will aid in making an enhancement selection. Chapter 5, “Case Studies and Future Applications” describes some examples of technology already being used in military operations.

3.1.2.2 Speech Under “Stress” Technology

3.1.2.2.1 Introduction

Military operations are often conducted under conditions of stress induced by high workload, high emotional tension, G-forces, and other conditions commonly encountered in the battlefield. These conditions are known to affect human speech characteristics. As a result, battlefield operations put military personnel under stress, which in turn impacts on their speech and the communications they have with speech processing equipment (such as voice coders, automatic speech (word) and speaker recognition) and other military personnel.

Speech under stress is defined here as speech that has had its characteristics modified as the result of an environmental (situation and/or physical) force applied to the person speaking. The applied forces (stressors) can be partitioned into four categories – Physical, Physiological, Perceptual and/or Psychological. Table 3.4 shows example stressors for each of these categories. See NATO Report #RTO-TR-10 dated March 2000.

Table 3.4: Examples of Stressors that Impact on the Production of Human Speech

	Categories of Applied Force			
	Physical	**Physiological**	**Perceptual**	**Psychological**
Examples of Stressors	- Acceleration (G-Force) - Vibration - Pressure Breathing - Breathing gas mixture - Speech Impediment	- Medication - Illness - Fatigue - Narcotics	- Noise - Hearing Defects - Poor grasp of language	- Workload - Emotional State such as anxiety and fear - Deception

Physical stressors are common in military environments. For example, there are multiple physical stressors to deal with in a fighter aircraft and they have large impacts on the speech produced by the pilot. Many of these stressors occur simultaneously, such as G-Force, vibration, cabin pressure, oxygen mask, to name a few. Some of these same stressors are found in tanks, underwater vehicles, as well as other types of military operations.

Physiological stressors include a wide range of stimuli; such as sleep deprivation, illness and medication that may produce second and/or third order effects. Perceptual stressors are common in military operations with the most common being the Lombard effect caused by noise. When an operator is subjected to high noise, he modifies the way he speaks to compensate for the noise. This perceptual effect produces Lombard speech, which has caused speech recognition systems to perform poorly. Perceptual stressors also have third order effects, such as the frustration that can be caused by difficulty in communicating over poor communication channels.

The range of psychological stressors is almost unlimited and the impact it has on the speech of the talker is highly individualistic as it is based on the individual's background and beliefs. For example, highly trained military personnel are mentally robust and react to fear and anxiety stressors differently than less trained civilian personnel. Researchers have observed that the change in speech produced by the trained military differs from that of the civilian.

These differences and the large number of stressors encountered in military operations impact on the technology used to reduce the effects produced on speech communications and speech recognizers. In fact, there are so many stressors and varied ways the stressors impact on the production of human speech that an array of technologies is required to reduce or compensate for these effects.

3.1.3 Stress Reduction Techniques

There are three methods for reducing the effects of stress on speech:

1) Equalisation or normalisation that attempts to return the speech features to their original characteristics;

2) Robust feature set, which depends on finding features that are robust to stress;

3) Model adjustment/training, which attempts to build models that incorporate the feature changes caused by stress.

The method used should be determined by the application and the stressors involved. Since the stressors influence the speaking style, stress reduction techniques or algorithms can and are designed around different speaking styles. For example, the Lombard speaking style discussed previously is produced by a noise stressor and techniques have been developed to lessen its impact on automatic speech technology. Other speaking styles include slow, fast, soft, loud, and angry speech. Each speaking style produces speech features that vary from normal or neutral speech. The effects of those speaking styles on automatic speech recognition are shown in Figures 3.5 and 3.6.

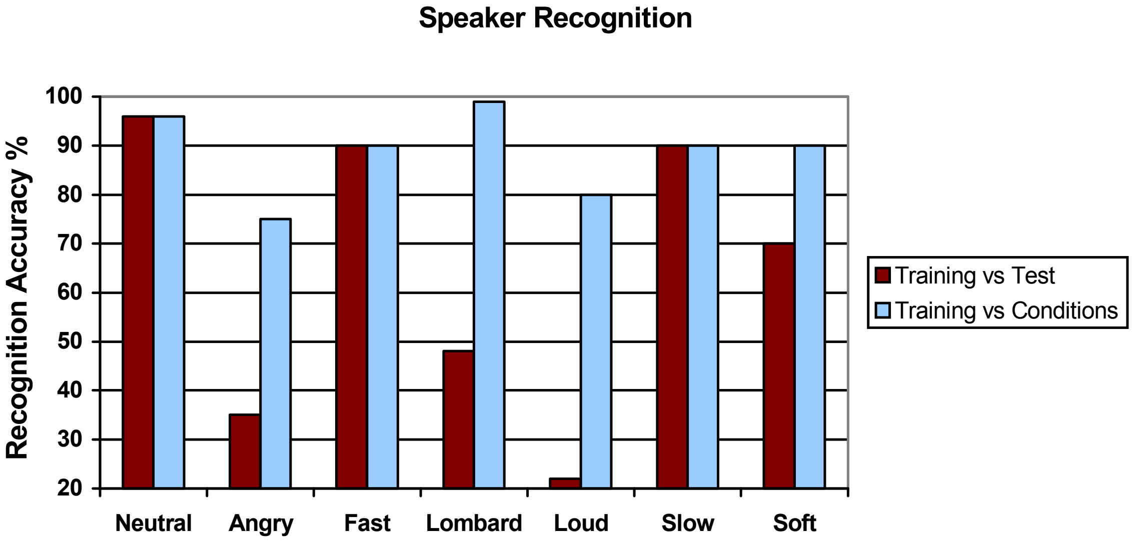


Figure 3.5: The Effects of Stress on Automatic Speaker Recognition.

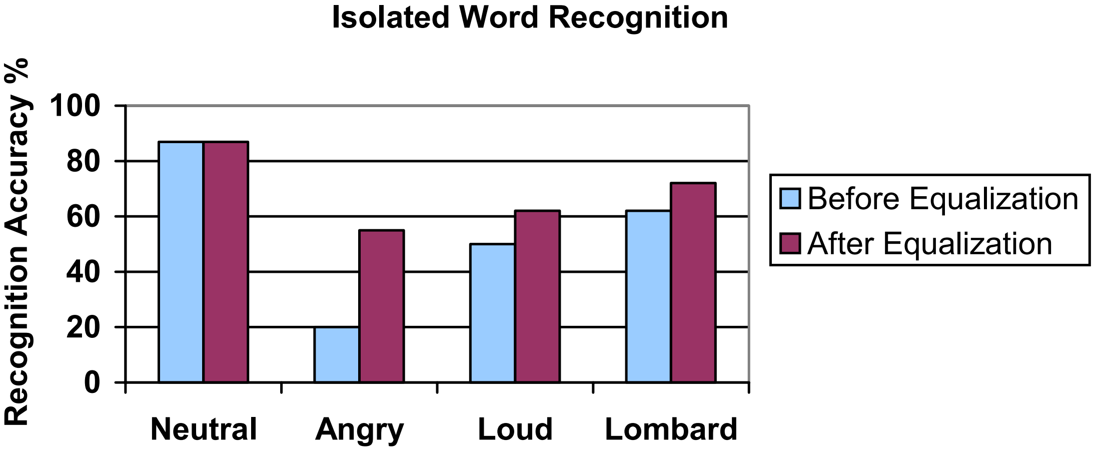


Figure 3.6: The Effects of Stress on Automatic Speech Recognition and the Improvement Obtained Using an Equalization Process on a 30 Word Vocabulary.

Figure 3.5 shows the results obtained for a speaker identification process for several speaking styles. When the process is trained on neutral speech, performance is seen to drop off quickly for the other speaking styles, especially loud and angry speech. Training for each speaking style gives good results, but is in most cases not practical operationally. Figure 3.6 shows the results obtained for an isolated word recognition process using a 30-word vocabulary with good SNR. The results show a decrease in performance from the neutral speaking style with angry speech producing the lowest score (20%).

The figure also shows the improvement that can be obtained using the robust feature method for reducing the effects of stress. This method is a laboratory model still in the development stage.

Of the three stress reduction methods, the model adjustment/training method is most often used even though it is difficult to use in operations that have highly variable stressors. Recently, work has been done in stress classification, that is, identifying the speaking style. Identification or classification is considered the first step toward improving the recognition accuracy of speech processing systems when the input is speech under stress. The state-of-the-art in stress classification is shown in figure 3.7.

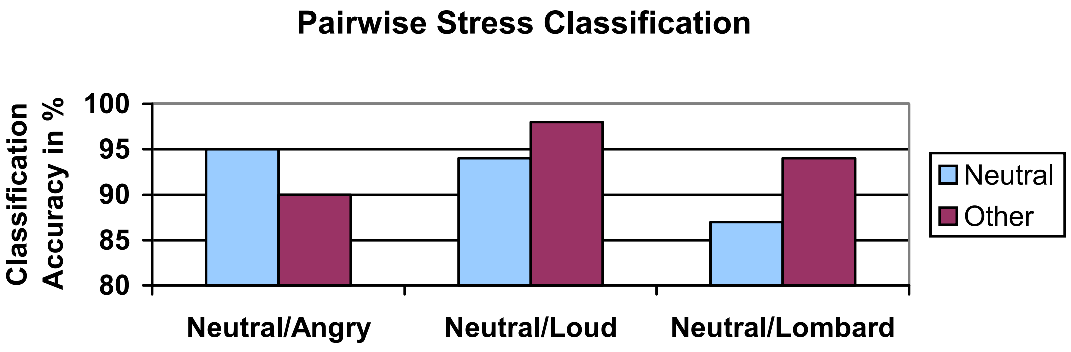


Figure 3.7: Pairwise Stress Classification using the Teager Energy Operator.

By using a classifier to identify the changes in speech and where they occur, appropriate processes can be applied to reduce the impact on voice communications equipment and military personnel.

An area of speech under stress that has been of interest to both military and law enforcement is the detection of deception. Interrogators and interviewers are interested in determining whether a subject is making a truthful statement or lying. The technology often called Voice Stress Analysis (VSA) is available commercially. The basic assumption underlying the operation of these commercial systems is the belief that involuntary detectable changes in the voice characteristics of a speaker take place when the speaker is stressed during an act of deception. The systems in general detect inaudible and involuntary frequency modulations in the 8-12 Hz region. The frequency modulations, whose strength and pattern are inversely related to the degree of stress in a speaker, are believed to be the result of physiological tremor or microtremor (Lippoid, 1971) that accompanies voluntary contraction of the striated muscles involved in vocalization. The systems generally use filtering and discrimination techniques and display the result on a chart recorder. Through visual examination of the chart by a trained examiner, a determination can be made on the degree of stress contained within a selected voice sample. The examiner looks for characteristic shapes related to amplitude, cyclic change, leading edge slope, and square waveform shapes called shapes blocking. Its success in detecting the stress in speech produced by deception is highly dependent on the examiner's ability to conduct the appropriate dialog, interpret the VSA chart, have incite of the physical and mental state of the subject, and on the subject's ability to hide the deception.

3.1.4 Speech Analysis

Speech analysis is the analysis of speech signals using digital signal processing techniques, independently of more high-level, linguistically oriented processing. One military relevant application is voice stress analysis.

3.1.4.1 Voice Stress Analysis

Voice Stress Analysers (VSA) are available commercially in two forms. One form used by many VSA vendors provides the capability on a laptop or PC with specific software packages with various levels of performance. Other vendors sell their product as an electronic box. Table 3.5 gives a list of vendors and their products. The information for this table was obtained from published literature and web sites listed under References.

Table 3.5: Table of Voice Stress Analysis Products

Product	Electronic Box	Computer Software	Manufacturer
PSE*	√		Dektor Counter Intelligence and Security
Lantern		Windows 3.11 & 95	The Diogenes Group, Inc.
Vericator		Windows 95, 98, NT 4.0	Trustech Ltd.
CVSA**	√		National Institute for Truth Verification
VSA Mark 1000 ***	√		CCS International, Inc.
VSA-15 ***	Hand-held		CCS International, Inc.
XVA 250	√		Xandi Electronics
TVSA3		Freeware	
Notes: * PSE – Psychological Stress Evaluator ** CVSA – Computerized Voice Stress Analyzer *** VSA – Voice Stress Analyzer			

The Lantern equipment manufactured by Diogenes Group Inc. consists of an analog magnetic tape recorder with an integrated microphone, a Pentium laptop computer, and proprietary software designed to run under *Windows* 3.11™ or *Windows* 95™**.** The recorder provides a stored record of all audio, while its monitor output provides a real-time input to a digital process that displays changes in spectral envelope. (These envelope changes are claimed to be related to microtremors caused by stress induced involuntary vocal tract muscle response.) Figures 3.8 and 3.9 show output waveforms that are interpreted by the user as a "No Stress" or "Stressed" voice pattern from the subject. In Figure 3.9, the subject shows stress caused by the subject's fear of spiders.

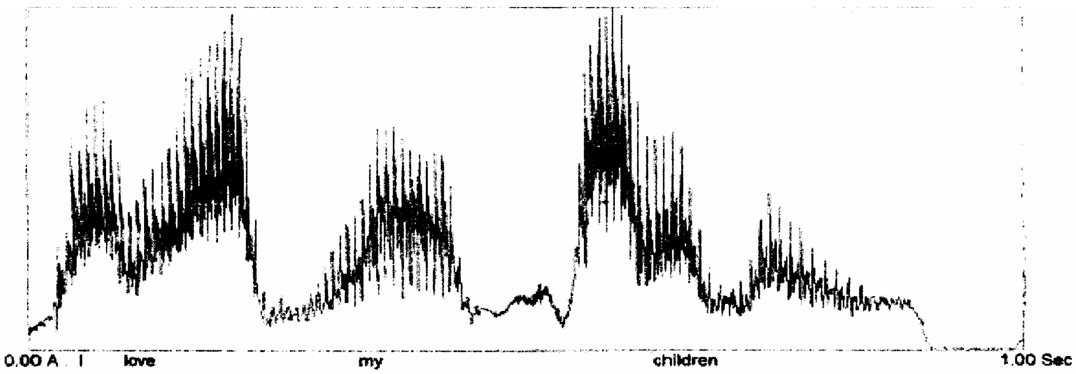


Figure 3.8: Diogenes Lantern Output Indicating No Stress.

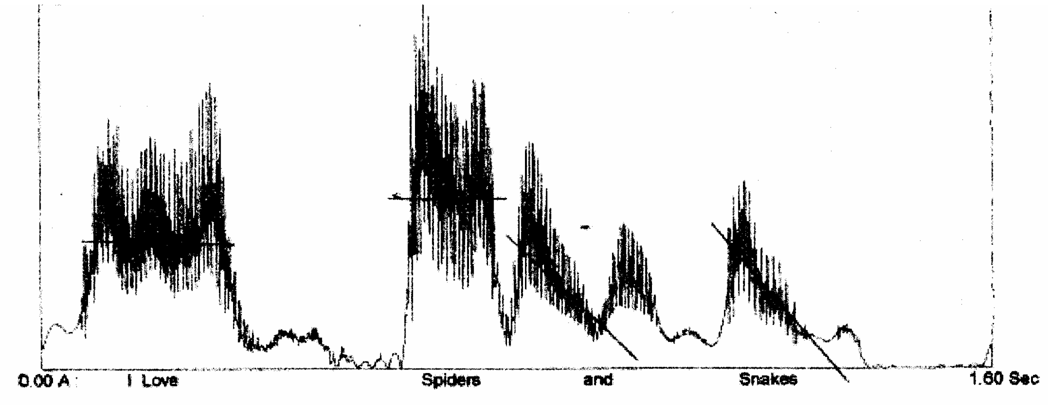


Figure 3.9: Diogenes Lantern Output Indicating Voice Stress.

The Vericator previously called Truster Pro is a software package that runs on a personal computer under *WIN95*™/ *WIN98*™/*NT* 4.0. The package includes a Vericator CD, Stereo T-connector and a user's manual. Truster, displayed on a monitor as shown in Figure 3.10 [Trustech LDT, 2001], is claimed to examine the emotional, cognitive and physiological patterns in the voice and provides a printed message as its output. Some of the messages outputted are: false statement, inaccuracy, deceitful, high stress, probably lying, and truth. In addition, the Vericator features automatic calibration, a graph display for advanced diagnosis, and a filtering process for reducing background noise.

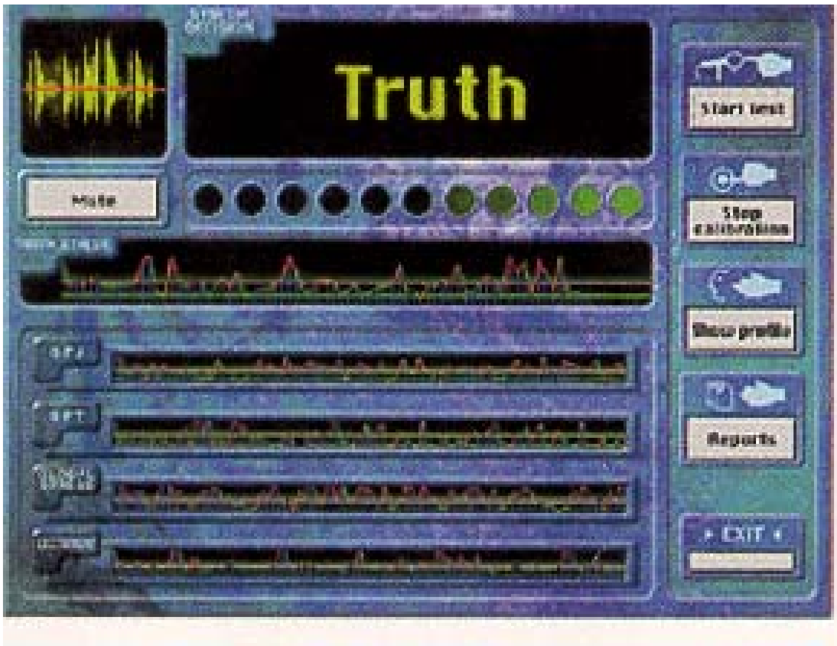


Figure 3.10: The Truster Stress Analyzer as Displayed on a Monitor.

The Lantern and Vericator products were evaluated in a recent test (completed in July 2001) [Haddad, 2001] that was conducted jointly by the Air Force Research Laboratory in Rome, NY, the National Institute of Justice, and local law enforcement agencies. One series of tests conducted was on field data to determine if these products could detect stress and potentially detect deception (lying). Two inactive murder cases were used for which ground truth was established through confessions, polygraph, and interviews. Each of the 48 utterances was analyzed by each system and compared to ground truth.

The Lantern indicated stress in its waveform plots in all utterances where stress indicators were justified by ground truth. The Vericator also scored 100% in indicating correctly some form of stress by displaying deceitful, high stress, or probably lying.

Another more subjective AFRL test [Haddad, 2001] was conducted in conjunction with two local law enforcement agencies over the last three years. Their assessment is that these products do detect stress, but stress does not always equate to a "lie". They believe that the voice stress products they tested are good interrogation tools and show their success rate in comparing test results with confessions to be in the high nineties.

Although these products have had success in interrogation scenarios, the success appears to be very limited in stored audio conversations. The reasons for this are: recordings are often of poor quality, conversational speech does not generally contain simple yes/no answers for analysis, and the number of stressors usually increase making it difficult to determine the cause of the stress.

3.1.5 Speech Synthesis

Speech synthesis is, in a very simplified definition, the capability of a machine to produce speech. How the task is performed, and what the system is capable of speaking is quite variable.

Techniques can be grouped into three main categories [Benoît 1995]: (1) the use of pre-stored digital messages; (2) concatenation of pre-recorded sentences or words generating new sentences that had not been entirely uttered before; (3) generation of speech from unlimited text using text-to-speech (TTS) synthesis. The third category is what usually is referred to as speech "synthesis" in the scientific community.

The first category is generally known as "canned speech", because the output speech is generated on the basis of pre-stored messages. The use of coding techniques to compress the message is also common in order to save storage space. Very high quality speech can be obtained, especially for quick response applications. This technique, however, requires large amounts of memory and is not very flexible. In the second category, a large number of new sentences can be generated from a limited set of pre-stored speech segments (e.g., a machine that reads bank account balances requires around one hundred segments to work properly). There are many existing applications of this technology in the telephone network. Drawbacks are the need of recording, labeling, and careful editing. This technique is not fitted for applications where there is a need for a large number of pre-stored segments, such as names. Naturalness can also be a problem if the segments are not carefully matched to the intended sentence in terms of prosody. The third category is the most sophisticated and will be described in more detail in next sections.

3.1.5.1 Text-to-Speech = Linguistic Processing + Speech Generation

Text-to-speech (TTS) synthesis allows the generation of any message from text. This generally involves a first stage of linguistic processing, in which the text-input string is converted into an internal representation of phoneme strings together with prosodic markers, and a second stage of sound generation on the basis of this internal representation. The sound generation can be made either entirely by rule, typically using complex models of the speech production mechanism, or by concatenating short pre-stored units.

3.1.5.2 Linguistic Processing

The first task in text analysis consists of dividing the text into words or similar units, and doing text normalisation. Normalisation is necessary, for example, to take care of numbers, converting them to dates, years and amounts. The system has also to deal with abbreviations, acronyms, etc.

It is well known that pronunciation of words generally differs from their spelling. After words have been identified, pronunciation can be found by looking them up in a pronunciation dictionary or by applying letter to sound rules. This results in a phonemic transcription of the input text.

The perceived naturalness of a TTS system is dependent on the richness of the intonative contour and rhythmic patterns, the so-called prosody. The main physical correlates are fundamental frequency, segment duration and energy. Conversion of text into prosodic parameters is essentially done in three phases: placement of boundaries, determination of duration and fundamental frequency contours.

3.1.5.3 Speech Generation

The last step for speech output is synthesis of the waveform according to the parameters defined at earlier stages of processing. Speech signal generators (the *synthesizers*) can be classified into three categories: (1) formant, (2) concatenative and (3) articulatory.

Formant synthesis, also referred as a rule-based approach, is a descriptive acoustic-phonetic approach to synthesis. Speech generation is performed by modeling the main acoustic features of the speech signal. The basic acoustic model is the source/filter model. The filter, described by a small set of *formants*, represents *articulation* in speech. Formants can be thought of as the frequencies at which vocal cavities resonate during articulation. The source represents *phonation*. Both source and filter are controlled by a set of phonetic rules (typically several hundred). High-quality rule-based formant synthesizers, including multilingual systems, have been marketed for many years.

In **concatenative synthesis** stored speech units, extracted from natural speech, are pieced together to produce speech. An example of the speech unit used is the phoneme, the smallest linguistic unit. Depending on the language there are about 35-50 phonemes in western European languages. The problem is combining the phonemes to produce natural sounding speech output. Fluent speech requires fluent transitions between the elements. A solution is to use diphones instead of phonemes. Instead of splitting at the transitions, the cut is done at the centre of the phonemes, leaving the transitions themselves intact. However, using this method the number of units increases to several hundred to cover all possible phoneme transitions in the language. Other units that are widely used are half-syllables, syllables, words, or combinations of them. To account for the large number of factors affecting properties of each speech segment, many different units are typically stored. A separate set of units must be collected and stored for every voice in each language. The cost of generating a corpus or an acoustic unit inventory is significant, besides making the speech recordings, each recording has to be analysed microscopically by hand to determine phoneme boundaries, phoneme labels and other tags. The synthesiser concatenates (coded) speech segments and performs some signal processing to smooth unit transitions and match prosody. Direct pitch-synchronous waveform processing is one of the most simple and popular concatenation synthesis algorithms. Several high-quality concatenative synthesisers, including multilingual systems, are marketed today.

While synthesizers based on diphone concatenation have proved the most popular in recent years, they are somewhat limited and difficult to improve upon. An alternate strategy is *unit selection,* where the second half of the synthesizer searches a large database of naturally spoken utterances and finds the best matching speech to the target. Usually this consists of a sequence of small speech fragments of one to three phones in length.

Articulatory synthesizers are physical models based on the detailed description of the physiology of speech production and on the physics of sound generation in the vocal apparatus. Typical parameters are the position and kinematics of articulators, such as lips and tongue. Then the sound radiated at the mouth is computed according to equations of physics. This type of synthesizer is far from use in applications because of its cost in terms of computation and the underlying theoretical and practical problems still unsolved. Even those that use concatenative synthesis, because it is currently the best available method, believe that in the long run that technique is not the answer [van Santen 1997, p.3]. "In the long term, articulatory synthesis has more potential ... for high-quality speech synthesis" [Shadle and Damper 2001].

3.1.5.4 Rule-Based versus Concatenative Approaches

There is little doubt that the better concatenation schemes exhibit superior voice quality. The rule-based approach has a clear advantage in the customisation it offers users for controlling speaking rate, pitch and voice characteristics. In addition, rule-based systems require dramatically less memory than concatenative ones, making them ideal for systems with memory limitations such as embedded systems. A detailed comparison can be found in [Hertz et al. 2000]. Development of rule-based systems requires a long period of trial and error analysis and human experts.

3.1.5.5 State of the Art

The current technology in text-to-speech synthesis already yields good intelligibility. However, it is still clearly inferior to human speech in terms of naturalness and the expression of emotion. In a recent evaluation [Juang 2001, p. 28], AT&T Next-Generation TTS system, based on unit selection, was rated for two different tasks at 3.46 and 3.91 in a 5 point voice quality and acceptability test (MOS). Intelligibility test scores were 3.48 and 3.98 in a five point scale, for the same system and tasks.

Speech synthesis is still an active research area. Some interesting developments are:

1) **More and more languages and voices** – as an example SpeechWorks provides TTS for US and UK English, Castilian and Mexican Spanish, Canadian and Standard French, German, Italian, Finnish, Brazilian Portuguese, Mandarin Chinese, Japanese, and Korean;

2) **Database methods –** increased use of statistical methods, supported by text and speech databases, for automatic word pronunciation, boundary placement, duration and fundamental frequency prediction. These methods are also applicable to automatic construction of the unit database used in concatenative systems;

3) **Creation of new voices by non-specialists** – several steps have been taken toward automatic inventory collection using real words, automatically selected, to record units, such as diphones, for concatenative synthesisers;

4) **Multilingual TTS systems –** using common algorithms for multiple languages [Sproat 1997]. Language specific information is stored in data tables;

5) **Voice Conversion –** methods for characterisation of a particular speaker, and conversion of the voice of a speaker into the voice of another speaker;

6) **Evaluation –** a major effort started in 1998 to make possible the evaluation of different systems with the same unknown material. 68 TTS systems for 18 languages participated in the first evaluation [Breen et al. 1998]. This process continued in the 2001 ISCA Tutorial and Research Workshop on *Speech Synthesis*;

7) **Combination of Natural Language Generation (NLG) with Text-to-Speech** – when the system generates the message it can also produce other useful information for the TTS system, avoiding the error prone analysis usually done in TTS when only text is available;

8) **Speech to speech –** Pioneering work in speech to speech technology has been done at Carnegie Mellon University beginning in the 1980's through Project JANUS. Several prototype applications

have been developed including a conversational speech translator for Spanish and a portable travel assistant. Another example is Verbmobil a speech-to-speech translation system for spontaneous dialogs in mobile situations. This multilingual system handles dialogs in three business-oriented domains, with context-sensitive translation between three languages (German, English, and Japanese);

9) **Visual Speech** – combining speech with synthetic faces or characters. Even auditory information is dominant in speech perception, it has been shown that visual cues increase speech intelligibility [van Santen 1997, p. 247].

Text-to-speech systems with good quality are presently available for a number of languages and require moderate computer resources (a normal PC equipped with an audio board is sufficient). It is clear that there is still a very long way to go before text-to-speech synthesis is fully acceptable, however, development is progressing.

3.1.5.6 Military Applications

Examples of present and future applications for speech synthesis in the military area are:

1) Voice warnings in aircraft cockpits as a complement for visual display of information. Speech synthesis can be used to transmit important information to pilots that can be missed in the middle of other displayed information;

2) Training of air traffic controllers, enabling training systems to simulate verbal communications with pilots;

3) Direction assistance via voice, for route-finding and situation awareness. As an example, a wearable speech enabled computer can give information to troops on the ground using stored maps, GPS and textual information from command stations, keeping eyes and hands free;

4) Combined with speech recognition and translation techniques, provide automatic language translation during international missions. Speech to speech takes speech processing a step further towards a voice-only computer interface. Several urban police departments in the US are currently experimenting with hand-held computers which can perform speech to speech in several different languages;

5) Text requires only a few hundred bits per second at normal speaking rates (and even less with proper coding). Thus text-to-speech synthesis, in connection with a reliable speech recognizer, would allow the ultimate in bit compression [Schroeder 1999].

3.1.6 Speech Recognition

Automatic speech recognition (ASR) is the capability of a machine to convert spoken language to recognized speech units, for example a word or phrase. These speech units are processed by an understanding stage, resulting in one or more actions. The action, which is a function of the application, could be for example the tuning of a radio receiver, a request for information or the conversion of a spoken input to text. Whatever the action, ASR can be valuable where the user’s hand and/or eyes are busy performing other tasks, such as a pilot flying an aircraft.

Currently, complex analysis methods, taking advantage of the increasingly available computer power, can be computed in real-time for limited domain dialog or in batch processing for large quantities of general domain, unrestricted speech.

3.1.6.1 Speaker Dependent or Independent Recognition

Speech recognizers can be speaker dependent or independent. Speaker dependent recognizers are trained by the person who will use the system. Although the recognition accuracy is generally better than it is for speaker independent recognizers, the user is burdened with the training process (that is, providing multiple spoken examples of speech units, words and/or phrases needed to perform the desired actions). Training may be particularly difficult for the user when the vocabulary becomes hundreds of words. Speaker dependent recognizers that automatically adapt to the speaker characteristics reduce the training requirements during use. Such systems are often called speaker adaptive systems and are delivered to the user with a factory-trained vocabulary.

Speaker independent recognizers attempt to maintain recognition accuracy independent of the person using the system. An advantage of speaker independent recognition is that no training by the user is required. A disadvantage is that recognition accuracy is generally less than it is for speaker dependent recognizers for the same vocabulary size.

3.1.6.2 Vocabulary Size

Classically, there has been a distinction in automatic speech recognition between “small vocabulary” and “large vocabulary” systems. The vocabulary of an automatic speech recognition system is defined as the set of words that the system is capable of recognizing. For the purpose of this paper we will define a small vocabulary ASR as a system for which all words in the vocabulary must be trained at least once.

In large vocabulary systems the recognition process starts by recognizing the speech sounds in the input signal and the sequences of sounds are then recognized as words. The latter step is generally carried out by looking up the pronunciation of words in a pronunciation dictionary, also known as the lexicon. This lexicon must have at least one entry for each of the words in the vocabulary. Assembling a lexicon is a laborious effort, and must be performed very carefully, as the performance of the system depends strongly on the quality of this lexicon.

The most direct application of a large vocabulary ASR is that of dictation. For this purpose, the vocabulary must be large enough to cover most of the words that will be used in the dictation task. If the domain is not exactly known, then the vocabulary must cover almost an entire language. For English, a lexicon of 20,000 of the most common words provides coverage on the order of 97.5% of the words used in newspaper texts. The comparable coverage for French is 94.6%, Dutch 92.8% and German 90%. For certain domains, such as the particular tasks in the military, the number of words needed in the vocabulary can be less than 1000 words, due to formal or procedural use of language.

At present there are good data resources in a number of languages, including English, Arabic, Japanese, German, Mandarin Chinese and Spanish. In English, there are hundreds of hours of speech available for thousands of speakers. There are two organizations that provide access to speech databases: LDC (Linguistic Data Consortium) in based North America and ELRA (European Language Resources Association) in Europe. Text corpora sizes for building language models are now exceeding 500 million words in English. Smaller data sets are available for many other languages. The advent of on-line newspapers has greatly increased the data available for this purpose.

If all possible speech sounds in a large vocabulary ASR have to be trained for the individual phones (typically 35-50, depending on the language), this training can become very tedious for an individual user. Therefore ASR systems can come with pre-trained models, that can be adapted to fit the features of a specific user.

3.1.6.3 Large Vocabulary Continuous Speech Recognition

Large vocabulary continuous speech recognition systems are normally speaker independent systems. Because of the large variability in speech produced by different speakers, these ASR systems have to be trained with many examples of speech uttered by many different speakers. These systems have to contend with a number of problems that are of less concern with simpler systems. The speech is largely unconstrained, resulting in the need for a large lexicon and intelligent software in order to determine where the word boundaries occur. They also have to contend with speakers using a variety of speaking styles and accents. In the case of broadcast news, the systems have to differentiate between speech, music, advertising jingles, etc. as well as constantly changing speakers.

Over the last 10 to 15 years there has been an increase in the difficulty of tasks used in evaluation of large vocabulary continuous speech recognition systems. In the early nineties, most of the development in large vocabulary speech recognition was made in (American) English. The US organization DARPA (Defense Advanced Research Projects Agency) has played a stimulating role in this development. ARPA not only sponsored speech recognition laboratories in their development of new systems and algorithms, they also organize a competitive assessment on a yearly basis, in which the performance of the various ASR systems is compared by an independent organization. DARPA sponsors the very labor-intensive task of recording acoustic training data bases, and makes sure that all training material used in the benchmark test are available for all competitors. These yearly assessment periods, together with all efforts in data collection, has had a tremendous influence on the performance of the ASR systems. In the 1995-2000 period, there was significant R&D and evaluations in two areas in particular: conversational telephone speech and broadcast news. The following are some recent performance results of ASR systems in these areas.

Table 3.6: Large Vocabulary Continuous Speech Recognition System Performance [Fiscus et al, 2000, Pallett et al, 1999]

Application Area	Language	Performance
Conversational telephone speech between friends	English	41.4% Word error rate
Conversational telephone speech between friends	Mandarin	57.5% Character error rate
Conversational telephone speech between strangers	English	19.3% Word error rate
Broadcast news	English	13.5% Word error rate
Broadcast news	Spanish	21.5% Word error rate
Broadcast news	Mandarin	20.6% Word error rate

Recently there has been considerable progress in commercialization of two applications of large vocabulary continuous speech recognition systems. The first and most universal are dictation systems. These are available from a variety of software companies including IBM (Via Voice) and Scansoft formerly Lernout and Houspie (Dragon Naturally Speaking). Dictation systems typically achieve word error rates of 5% or less, depending on the user. The second application is for automated telephone attendant systems. These systems replace the common button press menu with voice input to direct the user's call.

It is interesting to note that recognition performance is comparable to human performance. In a pilot experiment conducted in the European project SQALE [van Leeuwen et al, 1995], human listeners were offered sentences very similar to the DARPA “Wall Street Journal” tests, and they were asked to type in what they heard. For non-native English speakers (Dutch students) the average word error rate was 7%, while for natives (British and US students) this was 2.6%. Results from tests of ASR systems using a DARPA database of sentences read from Wall Street Journal and other North American Business papers gave word error rates of 7.2%.

Although the current state-of-the-art is very impressive, there are still many practical problems to overcome. Under adverse conditions, such as environmental noise, disfluencies, stress, cross talk, “incorrect” use of grammar, the word error rates increase. There is also considerable interest in recognition of previously encoded speech such Voice over Internet Protocol (VoIP). Talker variability is of major interest when dealing with non-native speakers and regional accents and dialects.

3.1.6.4 Robust Recognition for Adverse Conditions

The speech signal can be distorted in many situations by factors which impact on the quality of the speech produced by the speaker or by the transmission equipment or channel between the speaker and the recognition system. Some well-known distortions introduced by the speaker are the Lombard effect (the increase in vocal effort when a speaker is in a noisy environment), speaker changes caused by sickness, whispering, high g-load, etc. Distortions introduced by the environment and system are: noise, band-pass limiting, echoes, reverberation, non-linear transfer, etc. Other causes of deterioration of the speech signal may be the use of an oxygen mask in an aircraft, co-channel interference, jamming, etc.

Humans can recognize speech under very poor signal-to-noise conditions (down to -15 dB, depending on the noise spectrum). In these extreme conditions only digits and the alphabet (alpha, bravo, etc.) can be understood correctly. A speech recognizer cannot handle such adverse conditions. In general, the recognition performance decreases with signal-to-noise ratio. Optimal system development therefore requires an advanced signal treatment before recognition (i.e. noise canceling microphones, speech enhancement units) or compensation mechanisms within the recognizer.

For automatic speech recognition we may reduce the requirements for the recognition procedure in order to gain performance under these limited conditions. For example speaker independent large vocabulary recognition requires a speech signal of high quality while speaker dependent small vocabulary recognition for isolated words is more robust with respect to deteriorated signals.

As an example, the results of an isolated-word recognition experiment are given in Table 3.7. The test vocabulary included 25 command and control application words used in an advanced aircraft cockpit. Several conditions were investigated with and without the use of an oxygen mask and with a variety of added noise levels. For the condition without the oxygen mask, the speech was recorded in a noisy environment (noise level of 80 dBA) and with noise added afterwards, i.e. the speaker was speaking in a silent environment. This procedure was repeated for the condition where the speaker was using an oxygen mask. A specific noise-canceling microphone was located inside the mask (different from the standard mask mike). As the sound attenuation of the mask is significant, the noise level was increased to 100 dBA and 110 dBA which represents the noise level of a moderate fighter cockpit. The recognition scores (percent correct) are given in Table 3.7.

Table 3.7: Mean Recognition Scores (mean % correct) and Standard Errors (se %) based on One Speaker, Ten Test Runs, Seven Noise Conditions and With and Without the Oxygen Mask. The noise was introduced in two ways, directly with the speaker placed in a noise environment (direct) and by addition after the recordings (additive).

Scores in % correct		Without oxygen mask			With oxygen mask			
		no noise	direct 80 dBA	add. 80 dBA	no noise	direct 100 dBA	add. 100 dBA	add. 110 dBA
	mean	100.0	98.0	84.0	93.6	83.6	86.6	61.2
	se	0.00	2.0	5.9	0.7	3.1	2.7	3.6

The results show that the addition of noise has a significant effect on the performance of the recognizer. Also the effect of the oxygen mask is significant. A similar performance was obtained for the additive noise condition 80 dBA without oxygen mask and 100 dBA with oxygen mask. One could say that the use of the oxygen mask improves performance in the noise conditions due to the attenuation of the environmental noise. Optimal system design requires interaction between the applied front-end interfaces (e.g. microphone, signal processing) and the recognizer.

Tests have been conducted in aircraft cockpit environments using commercial off the shelf (COTS) products. In 1996 a recognizer produced by the Canadian Marconi Company was tested in a Bell 412 helicopter [Swail et al, 1997]. The small vocabulary system was used to control radio frequencies through a control display unit. The system was tested in actual flight conditions using simulated exercises. The overall results are shown in Table 3.8. In 1999 tests were conducted using the same Bell 412 helicopter and a Verbex Speech Commander system (Swail, 2001). Results of these tests are shown in Table 3.9.

Table 3.8: Word Error Rates for Test of Canadian Marconi Company Recognizer in Bell 412 Helicopter in Three Flight Exercises

Condition	Word Error Rate
Radio Exercise	5.2%
Simulated Mission	5.5%
Freeform Exercise	3.3%
Overall	5.1%

Table 3.9: Word Error Rates for Test of Verbex Speech Commander in Bell 412 Helicopter in Three Flight Conditions and a Quiet Hanger Test

Condition	Word Error Rate
Quiet (in hanger)	0.3%
Hover	3.1%
Cruise	2.8%
Confined Area Hover	3.9%
Overall	2.5%

3.1.7 Speaker Recognition

Speaker recognition is divided into two main categories. Speaker verification confirms the claimed identity of an individual, while speaker identification gives the identity of the speaker from the database of speakers the system knows about.

3.1.7.1 Speaker Verification

Speaker verification is a method of confirming that a speaker is the same person he or she claims to be. It can be used, often in conjunction with other means such as passwords or security cards, as a means of access control to secure systems. It provides an additional confirmation of the identity of the user that cannot be stolen. The heart of the speaker verification system is an algorithm running on a general-purpose computer or special-purpose signal processor that compares an utterance from the speaker with a model built from training utterances gathered from the authorized user during an enrollment phase. If the speech matches the model within some required tolerance threshold, the speaker is accepted as having the claimed identity. In order to protect against an intruder attempting to fool the system by making a recording of the voice of the authorized user, the verification system will usually prompt the speaker to say particular phrases, such as sequences of numbers that are selected to be different each time the user tries to gain entry. The speech verification system is combined with a recognition system to assure that the proper phrase was spoken. If it was not, the speaker will be rejected.

Depending on the application, different requirements can be imposed on the system based on the perceived costs of having the system make a mistake. A verification system can make two types of errors. It can reject a legitimate user (false rejection) or it can admit an impostor (false acceptance). Almost all systems allow for the adjustment of a tolerance threshold, making it possible to trade-off one type of error for the other. The overall performance of a verification system is generally expressed by means of its success rate SR=1-FAR-FRR, where FAR and FRR stand for the false acceptance and false rejection rates. Put in words, the SR represents the proportion of correct decisions issued on a given test set, summing up all accepted clients and rejected impostors.

There are several parameters that determine the difficulty of speaker verification for a particular problem:

- Quality of Speech. In the ideal application, the speech can be gathered in a quiet environment with a high-quality microphone and transmitted to the verification system over a wideband, high quality connection. This would be typical of a computer access control in which the user has direct access to the computer. In other cases, it may be necessary to access the system remotely by means of a telephone or other communications link. The collection and transmission of the speech may be significantly degraded making the verification problem much harder.
- Length of Utterance. Collecting a longer utterance will improve the performance up to a point, but will increase the time required for the test. There will be practical limits to what the user or the application will tolerate.
- Training Speech. Collecting more samples of speech from the user in multiple sessions over a longer period of time will improve performance by providing better models of the user. Over longer intervals it will also be necessary to update the training models as the speaker's characteristics change.

There are a number of factors that should be considered in comparing systems based on different algorithms. These factors include computational requirements of the algorithm and the sensitivity of the algorithm performance to differences between training and test conditions. The changes may result from changes in the speaker, the environment, or the channel. Algorithms for speaker verification generally match the characteristics of the speech of the user with the speech collected during training. Often some normalization is done to compensate for changing conditions on the channel or in the acoustic environment of the speaker. Gish and Schmidt (1994) and Campbell (1997) have prepared a survey of techniques.

Speaker verification is still an active area of research. Much progress has been made in the last few years. The US Department of Defense has developed a standard test corpus called YOHO that is designed for evaluating speaker verification systems. For these tests the performance standards were set as shown in Table 3.10.

Table 3.10: Speaker Verification Performance Specification

		Requirement	Goal
Probability	(False Rejection)	0.01	0.001
	(False Acceptance)	0.001	0.0001

In first tests with this corpus, a few systems were able to meet the minimum requirements, but none has yet achieved the long-term performance goal [Campbell, 1995].

3.1.7.2 Speaker Detection

The speaker detection problem is conceptually similar to the speaker authentication problem. The task is to sort a set of speech utterances by the likelihood that each is spoken by a particular speaker of interest. Potential users of such systems include law-enforcement and military authorities who may be searching for a speech messages spoken by a specific speaker amidst a vast number of messages spoken by other, irrelevant speakers.

Since 1997, NIST in the US has sponsored annual, formal evaluations of speaker detection systems [Doddington et al, 2000]. Performance is reported for each system in terms of probability of miss (similar to false rejection in the verification task) vs. probability of false alarm (similar to false acceptance in the verification task). Because a given system can be run at a variety of operating points, performance is often reported in terms of single operating point that permits easy comparison of systems, e.g. the equal error rate, the point at which the two error rates are equal. Given two minutes of training speech per speaker, and 30 seconds unknown test utterances, the best speaker detection systems exhibit equal error rates of approximately 8% on telephone speech. Factors leading to higher equal error rates include reducing the amount of training data, reducing the test utterance duration, imposing mismatched training and testing conditions (e.g. different handsets, background environments, channels), coding the training and/or testing speech (e.g. through VoIP) and so on. The best performing speaker detection systems at present employ Gaussian mixture modelling and universal background modelling [Reynolds, 1997].

3.1.7.3 Other Applications

Besides speaker authentication and detection, there are other related applications of speaker recognition technology. *Closed-set speaker identification* is used in applications where all the speakers to be encountered during recognition can each be enrolled during testing. Each test speech utterance is compared to each of the training models, and the speaker corresponding to the most likely model is hypothesized as the speaker of the test utterance. *Speaker tracking* is the process of determining those regions of a long audio cut during which a speaker of interest is speaking. *Speaker segmentation* is used to segment a long, multi-speaker speech cut into smaller fragments, each of which is spoken by a single speaker. *Multi-speaker detection* (or *open-set speaker identification*) is used in applications where one first finds messages spoken by any of a number of speakers of interest, and then further identifies which of the speakers of interest is speaking.

3.1.8 Language Identification

A language identification system is used to identify the language of the speech utterance. Automatic language identification systems generally work by exploiting the fact that languages have different phoneme inventories, phoneme frequencies, and phoneme sequences. These features can be obtained, although imperfectly, using the same spectral analysis techniques developed for speech recognition and speaker recognition. The use of higher-level features such as the prosodics and the use of expert knowledge about the target languages should also contribute to the language identification task, but to date, the best results have been obtained with systems which rely mainly on statistical analysis of spectral features.

There are practical issues that must be considered in putting together a system for a real application. Performance is of course a primary concern. This must be weighed against issues such as system complexity, the difficulty in training the system, and the ease with which new languages can be added. For example, the type and amount of data required for training could be very important. Some systems can be trained given only examples of conversations in a given language. Others require detailed phonetically marked transcriptions in each language. The relative importance of these issues will differ depending on the constraints of the particular application. A survey article by Muthusamy (1994) describes many of the techniques.

Some of the factors that make the language identification problem easier or harder are the following:

- The quality of the speech and the channel over which it is received.
- The number of possible languages from which the system must choose.
- The length of the utterance on which the decision must be made.
- The amount of training data which is available for making the models for each language. Both total duration of the training speech and the number of different training speakers are important factors.
- The availability of transcripts of the training speech, text samples from the language, and phonetic dictionaries for the language to assist in the creation of the models for the language.

Language identification continues to be an area of research. A series of tests was coordinated by NIST in the mid 1990s comparing the performance of language identification systems on a standard set of test data. In 1996 the evaluation focused on long-distance, conversational speech. Forty, 30-minute conversations per language were available for training in each of 12 languages. 20 conversations per language were used for testing. Test utterance size varied from 3 seconds to 30 seconds. The best systems exhibited 25% closed-set error rates on the 12-alternative, forced-choice problem on 30-second utterances. Average error rates of 5% were measured for the language-pair (language A vs. language B) experiments. [Zissman, 1997].

Current areas of research include reducing the training requirements, reducing the size of test utterances, and reducing the computational complexity.

3.1.9 Accent Identification

Automatic accent identification systems can be used in surveillance for law enforcement or military transmissions. Some criminal departments in the world have dialect experts that could be assisted by these systems. Other applications include the selection of services using the first/mother language of a non-native speaker thereby improving automatic speech recognition performance.

Preliminary results have indicated a 15% drop in recognition performance for non-native speakers when compared to native speakers using a native speech recognition system [Teixeira, 1992]. Selecting acoustic and phonotactic models, adapted to a specific accent, can improve the recognition performance significantly. However, this means the development of as many accent specific recognizers as accents identified for a particular language. Use of an accent specific system gave improvements of up to 60% when compared to the use of a conventional recognizer for native speakers. Although similar improvements can be found by adding the representative accents in the acoustical recognizer training material, the modularity of the proposed approach provides important advantages. Specific accent recognizers, which were previously developed and tuned according to the best suitable methodologies, can simply be integrated together in order to cover a wider range of speaker's accents [Teixeira, 1993, 1996, 1997, 1998]. Moving further away from the speaker independent recognition paradigm, better results may be found with more general approaches such as speaker adaptation [Tomokiyo, 2001].

As has been proposed for language identification, an accent identification system may also consist of a set of parallel speech recognizers. The one with the most plausible recognition output identifies, at the same time, the most plausible speaker accent. The following table represents the confusion matrix obtained from one of this type of system trained and tested with European male speakers for 200 English words. The overall accent identification rate was 85.3% providing an overall word recognition rate of 86.8%. However the recognition rate for the native speakers (UK) was 98.9%. Some of the figures in the confusion matrix can raise interesting explanations. Consider, for example, the first row of figures. Danish speakers provided the bigger number of non-native utterances classified as native (7.6%). They were the most difficult to identify as Danish, mainly because 19.3% of their words got a higher score from the Spanish recognizer. It is interesting to note that Danish people are generally fluent with English. On the other hand, these speakers were from Jutland having a strong tendency to transform the alveolar fricatives /s/ into palatal fricatives (/sh/) which is also common among Spanish speakers.

Table 3.11: Confusion Matrix (%) for an Accent Identification System (Teixeira, 1998)

Accent	Danish	German	English	Spanish	Italian	Portuguese
Danish	63.6	1.0	7.6	19.3	2.0	6.6
German	0.0	99.6	0.0	0.0	0.0	0.4
English	2.2	0.6	88.1	5.0	0.3	3.9
Spanish	3.7	1.7	4.9	83.7	0.0	6.0
Italian	0.0	3.7	1.2	0.5	91.9	2.6
Portuguese	2.0	2.7	7.4	3.7	2.5	81.8

The accents to be identified should be as well defined as possible. This will guide the collection of a representative speech corpus for training and testing and for choosing suitable features and classification methods [Benarousse 2001]. A first basic distinction should be made between non-native speakers and dialects or language varieties. Dialect differences are often significant and speakers do not generally attempt to conform to a standard variant. Non-native speakers, on the other hand, show different degrees of reading and pronunciation competence [Mengel, 1993]. Their knowledge of the grapheme-to-phoneme conventions of the foreign language may vary a lot, as well as their ability to pronounce sounds which are not part of their native sound inventory [Trancoso, 1999]. These differences justify the distinction of dialect or variant identification from accent identification that is here strictly addressed for non-native speakers [Brosseau 1992, Zissman 1996].

Non-native populations of speakers can be categorized in two different scenarios. One scenario considers immigrants or refugees. These speakers are under a long-term influence to acquire vocabulary and fluency according to the language variety used by the local population. A second scenario considers occasional travelers, such as businessmen, tourists and military personnel. These speakers usually have had a limited exposure to a standard variety of a foreign language (usually English) that will be needed in relatively few, but sometimes crucial circumstances.

Speaker classification according to their first/mother language can also be used for selecting a recognizer. This was also one application area for language identification, when the speaker actually uses his first language. However, for language identification, there is a vast amount of knowledge available about each specific language (phoneme inventories, grammar, etc.) which can not be used in a straightforward manner for accent detection. This can be considered among the reasons why accent identification is generally considered a more difficult task than language identification.

Instead of finding some qualities in the non-native speech such as the effects of the mother language, one might be interested in classifying it according to some kind of measure of performance in relation to a standard pronunciation in the second language. Giving feedback on the degree of nativeness of a student's speech is an important aspect of language learning. Skilled native speakers can easily discriminate at least five different ordered scores for classifying a student's utterance – average correlation between raters was once measured as 0.8 [Franco, 2000a]. In computer-aided language learning, this task has been addressed by many studies focusing on the segmental assessment of the speech signal [Neumeyer, 1996, Franco, 1998, 2000ab]. Recently, several studies have used suprasegmental speech information for computer-assisted foreign language learning [Delmonte 2000]. Some of these systems were able to obtain an average correlation between human and machine scores similar to the one obtained between different human scorers [Teixeira 2000, 2001].

In a world where globalization is an inevitable reality, accent identification will become a more important aspect of research in speech processing. Other new related research issues such as dialect [Chengalvarayan 2001], language identification [Benarousse 2001, Wong 2001], speaker recognition [Zissman 2001] and adaptation [Tomokiyo 2001] can bring new approaches to this area.

3.2 LANGUAGE PROCESSING

Language processing includes a number of technologies that can be used on text as well as speech input. In this paper we will consider topic spotting, translation and understanding systems.

3.2.1 Topic Spotting

Topic spotting is a term used to describe a number of techniques used to monitor a stream of messages for those of particular interest. The messages may be speech or text messages. Typical applications include surveillance and screening of messages before referring to human operators. Closely related methods may be used for automatic literature searches, message prioritization and some simple 'natural language' database interfaces.

Topic spotting of speech messages may be based on whole words or phrases (word spotting), phoneme sequences or even acoustic feature vectors. Central to topic spotting is the concept of 'usefulness'. To be useful a feature (e.g. a word) must occur sufficiently often for reliable statistics to be gathered, and there must be significant differences in the distributions between the wanted and unwanted messages.

Simple topic spotting applications may be constructed using manually chosen key words, however more advanced applications generally process training data to automatically determine suitable features. Care must be taken to ensure the training data is representative of the target domain. Although systems for processing speech messages may be trained on transcriptions of the training data, better results are often obtained by processing the output of a suitable recognition system. The latter approach includes effects of detection errors and recognition difficulty. For speech based topic spotting, various recognition technologies may be used, ranging from large vocabulary recognition systems, phoneme transcribers or the use of special ‘general speech models’ or ‘babble models’ in conjunction with whole word recognition systems. While word or phrase based systems generally ignore details of the context of the features, some form of clustering and fuzzy matching techniques is often desirable in phoneme and acoustic feature based systems. The application, need for reconfigurability and available computer power may significantly constrain the techniques which may be employed. Often the systems need not operate in real time.

The choice of the unit used for detection will partially determine what distinguishes the types of message. When processing speech data the decision to base the system on words, phonemes or acoustic feature vectors will affect what the system is sensitive to. For example, a system based on phonemes may be sensitive to regional accents as well as certain words, while a word-based system is more likely to be sensitive only to message content. Which is more useful will depend on the exact details of the application. Generally, to build up reliable statistics, a number of different features may be searched for an overall score for the message, based on the combined score. For continuous data, e.g. detection of weather forecasts in radio broadcasts, statistics are generally based on the frequency of occurrence of the features within some time interval. There may be a delay between the start of the section of interest and the output of the system.

Although topic spotting generally makes a binary decision – the message either is or is not of interest, similar techniques may be used to classify messages into one of a number of categories. Performance of topic spotting systems are often described in terms of ROC (receiver operator curves) curves. These show detection probability as a function of false alarm rate.

3.2.2 Translation

Machine Translation (MT) versus Computer-Aided Translation (CAT)

Long before personal computers made their entrance into every company, office and home, the advent of the first large analogue computing systems seemed to open a brilliant future with respect to overcoming the language barrier. Computer-generated translation of texts from any language into any other would just be a matter of time – so, at least, people thought in the 1950s. This euphoric view, well understandable from the technical standpoint, was, however, pretty soon followed by considerable disillusion which could be summarized in a relatively simple formula: You feed the computer with a set of well-defined grammatical rules and provide it with a rather large vocabulary, but you cannot make the computer think like a human being and teach the machine the knowledge that in translation science is often referred to as “world knowledge” – a complex product of education and experience.

Skeptical attitudes towards machine translation reached a climax in 1966 when the US Automatic Language Processing Advisory Committee (ALPAC) published a report in which experts recommended to completely stop the financial support for MT research because the quality of the current MT results was simply too poor. Furthermore, they argued, the translation market would not grow significantly, there would be enough human translators to meet the demand and, above all, human translation would be much cheaper.

Since then, the situation has, of course, changed considerably, both on the hardware and software side. But nevertheless the state of the art of machine translation has in general not yet reached a quality level which can satisfy the translation requirements for all sorts of texts and language pairs. It is undisputed that MT is quite successful in domains where texts have a rather uncomplicated structure and syntax and where the content cannot be misinterpreted. But even in such cases, human intervention is necessary, by post-editing the machine translation output and/or by customizing the MT system and pre-editing the documents to be translated. MT results can further be enhanced, when the structure and style of the source documents is already influenced in statu nascendi, i.e. when they are written. For instance, some companies with in-house technical writers have the documentation (user handbooks, troubleshooting guides, online help texts, etc.) for their products written in a “controlled language”, i.e. the authors use a limited vocabulary and build straightforward sentences (avoiding long subordinate clauses, ambiguous formulations and relations, etc.). Texts in which style and subtleties of language play a major role (as in literature or poetry) will undoubtedly never be promising candidates for machine translation.

Around the beginning of the 1990s, the so-called computer-aided or computer-assisted translation systems (CAT systems) became commercially available. In contrast to MT, where no human interactivity is involved in the actual process of translating, CAT systems are designed as a software and/or software/ hardware combination acting as supplementary translation tools to increase the performance and productivity of a human translator. In CAT systems, the computer only takes over the tasks that a machine can do best, while the tasks that require human intelligence are left to the translator who remains in control of the overall translation process. CAT tools comprise a range of different products, from on-line dictionaries and spell checkers to translation memory technologies and terminology management systems.

Among the current CAT systems based on translation memories (TMs) are: CypreSoft, DéjàVu, IBM Translation Manager, SDLX, Trados Translator’s Workbench and Transit. The Translator’s Workbench developed by Trados has emerged as the leading TM system on the market. TM systems take advantage of the fact that in certain kinds of texts, especially in technical documentation, there are a considerable number of repetitive or similar sentences, either within the document itself or within a collection of documents (especially when dealing with updates, revisions, etc.). The TM algorithm not only automatically recognizes fully identical sentences but also variants with differing degrees of similarity. When working with a TM system, each sentence – or segment – that has been translated and validated by the translator is entered into a translation memory. This translation memory is a database of segment or sentence pairs where each source language segment is paired with its corresponding target language segment. So, when the same or a similar sentence has to be translated again, whether in the same document or a different one, the system automatically retrieves the previous translation from the memory. The sentence is shown either as a perfect (100 %) match when the segment found in the TM is identical, or as a fuzzy match when the segment found is not completely identical, but similar. (The degree of fuzziness up to which matches are retrieved is a user-defined percentage value; matches below this threshold value will not be displayed.) The translator can then accept the proposed translation, edit it and validate the final translation in order to store it in the translation memory.

Usually TM systems come along with a terminology management system that runs in the background of the translation process. All terms that occur in the current segment and are found in the terminology database are automatically recognized and displayed in a separate window from where the translator can copy them. In this manner, terminological consistency within one or more documents, or even a whole translation project, is ensured. Another component usually provided with a TM system is an alignment tool that enables the user to build a translation memory from previous translations (for which no TM system has been used). This is a very helpful function because the TM system ships with an empty memory. In order to benefit from the TM technology, it is of utmost importance to build a large and powerful translation memory as quickly as possible. The simple philosophy underlying all TM systems is: You can’t get anything out of the system that you have not put in before!

CAT systems provide a lot of other specific features whose explanation would go beyond the scope of this short presentation. For more information on the Trados Translator’s Workbench, readers should visit the website: http://www.trados.com.

Currently, many CAT systems are best described as “hybrid” systems as they offer an interface to an external MT system. The documents to be translated can be processed by the associated MT system either interactively (i.e. if desired, the user activates the MT link in order to get a translation proposal for the current segment) or in batch mode (i.e. the whole document is pre-processed by the MT system). Such hybrid approaches offer quite promising ways to combine CAT technologies with MT technologies.

It may well be that in a not too distant future the hitherto opposing concepts of machine translation and computer-aided translation will enter a close and beneficial symbiosis.

3.2.3 Understanding

Speech systems have progressed to the point at which they can begin to handle tasks that could broadly be described as language understanding. There is a tremendous range of problems that could fall under this definition. Systems are starting to be applied to real, practical problems at the simplest end of this range.

Understanding problems can be divided into two broad categories. The first set of problems addresses human-machine interactions. In this case the person and the machine are working jointly to solve some particular problem. The interactive nature of the task gives the machine a chance to respond with a question when it does not understand the intentions of the user. The user can then rephrase the query or command. In the second type of problem, the machine has to extract some desired information from the speech without the opportunity for feedback or interaction.

The best examples of the human-machine interaction are found in some simple information retrieval systems. Such systems are beginning to move from laboratory prototypes into field demonstrations. Recently a spoken language systems program has used an Air Travel Information System as a model problem to drive the development of the technology. In this system, the user makes voice queries on an actual airline reservation database in order to make a plane reservation. Similar systems are being explored for use with actual reservations systems.

Another example is the prototype voice-driven telephone system that is used by the German Railways to provide schedule information for trains between cities in Germany. The system has a vocabulary of about 2000 words including 1200 station names. In spite of a 25% word error rate, the system is able to provide the correct response to more than 80% of the queries with less than 2 corrections.

An example of an experimental system which does not permit feedback is one used to monitor air traffic control messages [Rohlicek et al, 1992]. This system tries to extract the flight identity from the voice message. Such a system has been considered as an aid to the air traffic controller. The system would identify the flight when a voice transmission is received and automatically highlight the information on the controller's display.

Further advances in understanding systems will build on progress in many fields, including speech recognition, natural language understanding and man-machine interaction.

3.3 INTERACTION

In this section we will cover a number of advanced topics in speech technology. They are interactive dialog, multi-modal communications, and 3-D sound.

3.3.1 Interactive Dialogue

A dialogue is usually considered to be an interchange between two cooperating partners during which some information is passed from one to the other. It may be better to treat the concept differently; recognizing that one of the partners has initiated the dialogue for a purpose that is not simply to conduct the dialogue. Accordingly, the two partners in a dialogue should be considered asymmetrically, one being the originator of the dialogue, the other being the recipient. The dialogue itself is successfully concluded when at least the originator believes that the recipient is in the intended state. The intended state may be that the recipient now has some information, or that the recipient has provided some information, or that the recipient is performing some task on behalf of the originator. In effect, a single one-way message has passed between originator and recipient, and has had a desired effect observable by the originator.

The back and forth flow associated with the common concept of "dialogue" consists of messages at a lower level of abstraction that enable the originator to ensure that the recipient receives the intended message (i.e. comes to a state that the originator intended). The dialogue "supports" the transmission of the one-way prime message. Each of these supporting messages, in turn, can be considered in the same way as the main message, as a one-way message that is to be supported by a two-way dialogue of messages at successively lower levels. What is commonly thought of as a "dialogue" can in this way be considered as a tree structure of messages in which each branch point represents a message in one direction or the other, and the leaves represent the physical realizations of the communication.

When the recipient is a computer, the same considerations apply as when the recipient is a human. The human originator wants to get the computer into some state, either to provide it with data, to get it to provide data, or to get it to perform some task. What the human wants the computer to do may not be something that can be done with one prepackaged command, and a dialogue is necessary. When the mode of the dialogue includes speech either as input to the computer or as output, or both, it is much less certain that the computer has correctly interpreted anything the human says than would be the case if the human entered the same words on a keyboard. Therefore there must be a possibility for a dialogue that supports the transmission of the individual utterances. Interpretation errors must be perceptible to the human as well as correctable when they occur.

Dialogue using speech I/O is not different in principle from dialogue with a computer using keyboard and display, but is valuable for different purposes and under different circumstances.

3.3.2 Multi-Modal Communication

Speech is a natural mode of communication for humans, but, having evolved over thousands of years, language has become very complex. Machines are evolving too, but in the beginning they could utilize only very simple interfaces, such as switches. As the capabilities of computers have increased, more complex modes of communication with them have become possible, including the keyboard, the light pen and the mouse. Recently, computers have evolved to the point where the use of speech and natural language interfaces are also possible. The situation now is that several methods, or media, are available for human-computer interaction: this is described as a multi-media interface.

At present, the various media in a multi-media interface operate in parallel or as alternatives. Each mode of communication has its own sphere of operation. For example, a mouse is ideal for moving the cursor to a particular place on the screen, but far from ideal for adding text at that point. The keyboard is very capable for adding text, but far from optimal for placing the cursor, especially if it has to be moved a long way. Considering a much more complicated example, a pilot could enter waypoint data into his navigation computer by voice, but would find switches more natural, or quicker, or more reliable, for other functions like lowering the undercarriage. Each interface mode is best suited to a particular kind of operation; matching the medium to the message is the essence of the art of interface design.

The next stage in this evolution is to make the various modalities interact and cooperate in carrying out tasks; this is the multi-modal interface. Now, each individual command to the system may utilize several interface modalities. This capability already exists in a limited form in the mouse, which combines pointing and a switch operation in a manner that seems natural to the user, or at least, is very easily to learn. The addition of speech to multi-modal interfaces will greatly increase their power. The combination of voice input and a pointing device would allow commands to use words like 'this', 'that' or 'there', perhaps allowing a simplified display and removing the need for the operator to remember details of unit designations, etc. The associated displays and auditory outputs would need to be an integral part of such an interface, each tailored to provide information to the operator in the most easily assimilated form.

In military systems, the problem that the multi-modal interface addresses is the mismatch between the increasing complexity of weapons systems and the more or less static capabilities of military personnel. The later can, of course, be improved by training, but this may be an expensive option, especially for those forces comprised partly of conscripts and thus having a rapid turnover of personnel. The alternative is to design the interface between the man and the machine in such a way that the machine comes to seem simple, or, at least, considerably less complex than it is. The combination of language (usually spoken but not always) and gesture is natural to humans and therefore minimizes workload. A command such as 'Put that over there', with the accompanying gestures, can be issued and understood by humans without significant effort.

The integration and fusion of information made possible by the multi-modal interface will allow the man-machine dialogue to be carried out at a higher level of abstraction than before. This is appropriate to the use of speech and language, as the words used may cover a wide range of levels of abstraction.

Multi-modal systems can also be used to improve the performance of a single mode system. We will consider the case of a speaker verification system. Compared to other verification systems, speaker verification benefits from only requiring relatively cheap microphones as input sensors. Systems that need dedicated hardware, like fingerprint analysis or iris recognition, often results in increased production costs. They also tend to suffer from relatively low user friendliness due to their close proximity constraint, but offer a high level of performance. On the other side, systems dealing with widely available multi-media sensors as those found around personal computers (microphones and cameras) are cheaper to produce, better accepted for their convenience, but cannot compete with the performance offered by dedicated hardware. To compensate for this lack of performance, multi-modality – i.e. using combination of different algorithms and feature sets – is often considered.

Under the framework of the M2VTS project – an EEC-funded project dealing with person authentication from multiple clues – several authentication algorithms based on different modalities, have been benchmarked using a common multi-modal database [Pigeon, 1997]. As far as single modalities were concerned, speech authentication offered the highest performance with a success rate of 97.5%. By combining speech information with labial features found in the associated video sequence, the SR increased up to 99.4% [Jourlin, 1997]. A face-based technique running on static frontal images, achieved a SR of 89% and profile information taken from a side-view, achieved the same level of performance. By combining frontal and speech features together, the SR rose to 99.5% [Duc, 1997]. Highest performance has been achieved when combining frontal, profile and speech features all together, with a SR of 99.95%.

These excellent authentication rates of over the 99% are boosted in fact by the high performance offered by the speech modality on its own. This illustrates the important role played by speech-based technology inside multi-modal systems.

In the area of speech recognition, research is being carried out to investigate the use of myo-electric signals generated by facial muscles to improve speech recognition [Chan, 2001]. Surface mounted electrodes were embedded in a pilot's oxygen mask to collect the facial muscle signals. Classification errors for just the myo-electric signals were found to be 2.7% and 10.4% in two recognition experiments carried out using a ten-digit vocabulary. The oxygen mask is a difficult acoustic environment for speech recognition and combining the acoustic signal with myo-electric signals shows potential to enhance recognition performance.

3.3.3 Dimensional Sound Display

Modern signal processing techniques allow headphone audio to be processed in such a way that it seems to originate from virtual sound sources located in the three-dimensional space around the listener. By using head-tracking devices, it is even possible to create a stable virtual acoustic environment that takes (head) movements of the listener into account. One application of such a 3D-auditory display is enhancement of situational awareness by using virtual sound sources to indicate positions of relevant objects (e.g. targets or threats). Bronkhorst et al. (1996) have shown in a flight simulation experiment that 3D-sound can indeed be used as an effective acoustic radar display. They found that the time required to locate and track a target was similar for acoustic and visual displays and that the combination of both displays yielded even shorter search times.

A second, perhaps less obvious, application of 3D-auditory displays is their use as a means of improving communication and signal detection. This application is based on the ability of the human hearing system to tune in on sounds coming from one direction while suppressing sounds coming from other directions, a phenomenon which was coined as the "cocktail party effect". The results of a number of identification experiments clearly demonstrate this effect. One target speaker and up to four interfering speakers were presented simultaneously in three headphone conditions: monaural, binaural without signal processing and binaural with 3D-sound. In the second condition, target and interfering speakers were divided over both ears; in the latter condition, the speakers were presented through virtual sources placed on a 180° arc in front of the listener. The results, obtained from six listeners, are plotted in Figure 3.11. There appears to be a clear superiority of 3D-sound presentation over the other presentation modes.

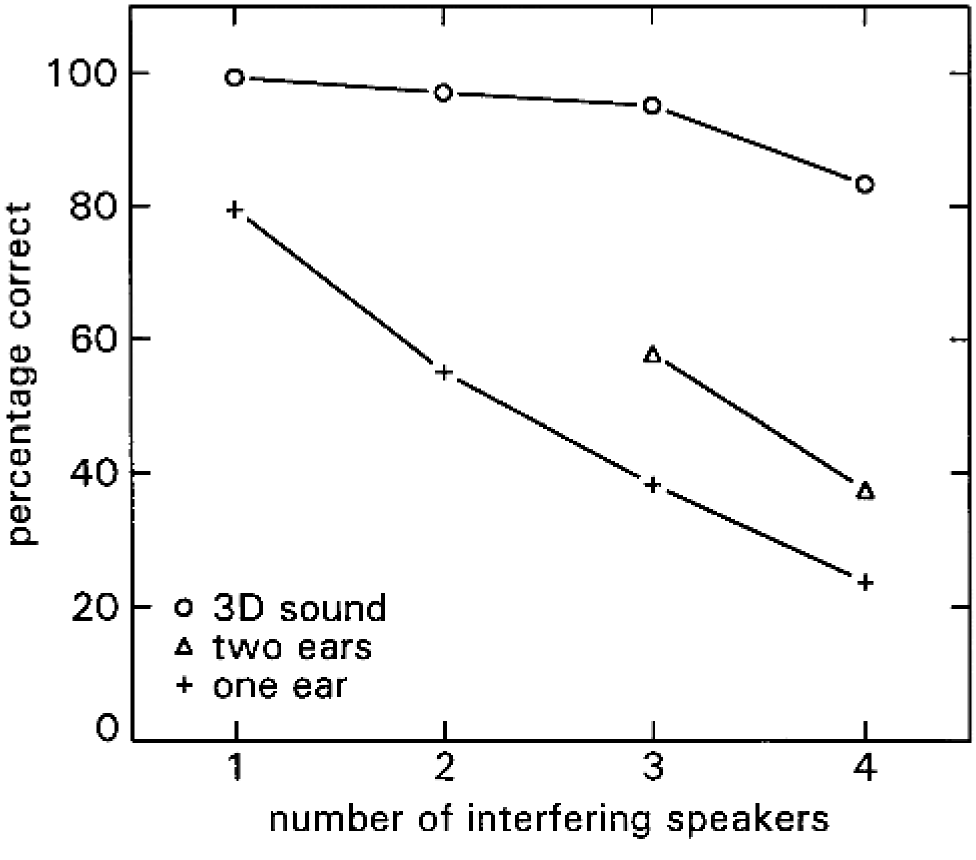


Figure 3.11: Results of a Number Identification Experiment, in which a 3D Auditory Display was Compared with Normal Sound Presentation.

A 3D-auditory display, therefore, offers significant advantages to persons who have to deal with multiple simultaneous signals. The "cocktail party effect" will help them to suppress unwanted sounds and the signals can also be presented from "meaningful" directions, to facilitate identification. Given the abundance of (military) functions where such situations occur (pilots, flight controllers, sonar operators, etc.), it seems probable that 3D auditory displays will become widely applied within the next few years.

3.4 REFERENCES

Arslan, L.M. (1996a). "Automatic Foreign Accent Classification in American English", PhD thesis, Duke University, Durham, North Carolina.

Arslan, L.M. and Hansen, J.H. (1996b). "Language accent classification in American English", Speech Communication, 18(4):353-367.

Benarousse, L., Geoffrois, E., Grieco, J., Series, R., Steeneken, H., Stumpf, H., Swail, C. and Thiel, D. "The NATO Native and Non-Native (N4) Speech Corpus", Proc. of the Workshop on Multilingual Speech and Language Processing, MSLP1, Aalborg, Denmark, September 2001.

Benoît, C. (1995*). Speech Synthesis: Present and Future*. in European Studies in Phonetics & Speech Communication, G. Bloothooft et al. (Eds). OTS Publs.

Berkling, K., Zissman, M., Vonwiller, J. and Cleirigh, C. (1998), "Improving Accent Identification Through Knowledge of English Syllable Structure", In Proc. Int. Conf. on Spoken Language Processing, Sydney, Australia.

Blackburn, C., Vonwiller, J. and King, R.W. (1993). "Automatic accent classification using artificial neural networks", In Proc. of the European Conf. on Speech Comm. and Tech., Volume 2, pp. 1241-1244, Berlim. ESCA.

Breenen et al (eds) (1998). Proceedings of the Third International Workshop on Speech Synthesis. Available at http://www.slt.atr.co.jp/cocosda/jenolan/index.html (checked 6 November 2001).

Bronkhorst, A.W., Veltman, J.A. and Breda, L. van (1996). "Application of a Three-Dimensional Auditory Display in a Flight Task". Human Factors, 38(1), pp. 23-33.

Brousseau, J. and Fox, S.A. "Dialect-dependent speech recognisers for Canadian and European French", Proc. ICSLP, Banff, 1992, Volume 2, pp. 1003-1006.

Campbell, J.P. "Speaker recognition: A tutorial," Proceedings of the IEEE, Vol. 85, pp. 1437-1462, September 1997.

CCS International, 4 October 2001, <http://www.spyzone.com/>

Chan, A.D.C., Englehart, K., Hudgins, B. and Lovely, D.F., "Myo-electric signals to augment speech recognition", Medical and Biological Engineering and Computing, Vol. 39, no. 4, 2001, pp. 500-504.

Chengalvarayan, R. "HMM-Based English Speech Recognizer for American, Australian, and British Accents", Proc. of the Workshop on Multilingual Speech and Language Processing, MSLP7, Aalborg, Denmark, September 2001.

Cole, R. (Editor) (1996). *Spoken Output Technologies*. In Survey of the State of the Art in Human Language Technology (1996), Ronald A. Cole, (Editor in Chief), chapter 5. http://cslu.cse.ogi.edu/HLTsurvey/ (checked 2 November 2001).

Decision Sheet-2, NATO C3 Board Communication Network Sub Committee Ad Hoc Working Group on Narrow Band Voice Coding, AC/322(SC/6-AHWG/3)DS/2.

Decision Sheet-5, NATO C3 Board Communication Network Sub Committee Ad Hoc Working Group on Narrow Band Voice Coding, AC/322(SC/6-AHWG/3)DS/5.

Decision Sheet-9, NATO C3 Board Communication Network Sub Committee Ad Hoc Working Group on Narrow Band Voice Coding, AC/322(SC/6-AHWG/3)DS/9.

Delmonte, R. "SLIM prosodic automatic tools for self-learning instruction", Speech Communication, Vol. 30, pp. 145-166, 2000.

Diogenes Company, 9 October 2001, <http://www.diogenesgroup.com/>

Doddington, G., Przybocki, M., Martin A. and Reynolds, D.A. "The NIST speaker recognition evaluation – overview, methodology, systems, results, perspective," Speech Communication, Vol. 31, pp. 225-254, March 2000.

Duc, B., Maître, G., Fischer, S. and Bigün, J. "Person Authentication by Fusing Face and Speech Information", Proc. First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA '97), Crans-Montana, Switzerland, March 12-14, 1997, pp. 311-318.

Dutoit, Thierry (1997), *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers (Dordrecht), ISBN 0-7923-4498-7, 312 pages.

Fiscus, Jonathan G., Fisher, William M., Martin, Alvin F., Przybocki, Mark A., Pallett David S. Proceedings of the 2000 Speech Transcription Workshop May 16-19, 2000, http://www.nist.gov/speech/publications/tw00/index.htm 2000 NIST EVALUATION OF CONVERSATIONAL SPEECH RECOGNITION OVER THE TELEPHONE: ENGLISH AND MANDARIN PERFORMANCE RESULTS, http://www.nist.gov/speech/publications/tw00/pdf/cts10.pdf.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R. and Cesari, F. "The SRI EduSpeak System: Recognition and pronunciation scoring for language learning," Proc. of Integrating Speech Technology in Language Learning, 2000a.

Franco, H., Neumeyer, L., Digalakis, V. and Ronen, O. "Combination of machine scores for automatic grading of pronunciation quality," Speech Communication, Vol. 30, pp. 121-130, 2000b.

Franco, H. and Neumeyer, L. "Calibration of machine scores for pronunciation grading", Proc. Int'l Conf. on Spoken Language Processing, 1998.

Fung, P. and Liu, K. (1999). "Fast Accent Identification and accented speech recognition", ICASSP, Phoenix, Arizona, USA, March 15-19, 1999.

Gish, H. and Schmidt, M. "Text-lndependent Speaker Identification", IEEE Signal Processing Magazine, October 1994, pp. 18-32.

Haddad, Darren, "Investigation and Evaluation of Voice Stress Analysis Technology," USAF/AFRL Technical Report, July 2001.

Hansen, J.H.L. and Arslan, L. (1995). "Foreign accent classification using source generator based prosodic features". In Proc. Int. Conf. on Acoustic Speech and Signal Processing, Volume 1, pp. 836-839, Detroit.

Hertz. S. et al. (2000). *Space, Speed, Quality, and Flexibility: Advantages of Rule-Based Speech Synthesis*. Proceedings of AVIOS 2000.

Huang, B.H. (Editor) (2001) Special Issue on Speech Synthesis. IEEE Transactions on Speech and Audio Processing. Vol. 9, No. 1, January.

Huang, X., Acero, A. and Hon, H. (2001). *Spoken Language Processing*. Chapters 14, 15 and 16. Prentice Hall.

Humphries, J. and Woodland, P. (1998). "The use of accent-specific pronunciation dictionaries in acoustic model training". In Proc. Int. Conf. on Acoustic Speech and Signal Processing, Seattle.

Jourlin, P., Luettin, J., Genoud, D. and Wassner, H. "Acoustic-Labial Speaker Verification", Pattern Recognition Letters, Vol. 18, No. 9, September 1997, pp. 853-858.

Kumpf, K. and King, R.W. (1997). "Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks". In Proc. of the European Conf. on Speech Comm. and Tech., pp. 2323-2326, Rhodes, Greece.

Lippold, O. "Physiological Tremor," Scientific American*,* Vol. 224, No. 3, pp. 65-73, March 1971.

Mengel, A. (1993). "Transcribing names --- a multiple choice task: mistakes, pitfalls and escape routes". In Proc. 1st ONOMASTICA Research Colloquium, pp. 5-9, London.

Muthusamy, Y., Barnard, E. and Cole, R. "Reviewing Automatic Language Identification", IEEE Signal Magazine, October 1994, pp. 33-41.

National Institute for Truth Verification, 4 October 2001,<http://www.cvsal.com/>

NATO STANAG 4198, "Parameters and coding characteristics that must be common to assure interoperability of 2400 bps linear predictive encoded digital speech".

NATO STANAG 4209, "The NATO Multi-Channel Tactical Digital Gateways Standards for Analogue to Digital Conversion of Speech Signals".

NATO STANAG 4479, "Parameters and coding characteristics that must be common to assure interoperability of 800 bps digital speech encoder/decoder".

Neumeyer, L., Franco, H., Digalakis, V. and Weintraub, M. "Automatic text-independent pronunciation scoring of foreign language student speech", Proc. Int'l Conf. on Spoken Language Processing, pp. 1457-1460, 1996.

Pallett, David S., Fiscus, Jonathan G., Garofolo, John S., Martin, Alvin and Przybocki, Mark. DARPA Broadcast News Workshop February 28-March 3, 1999, http://www.nist.gov/speech/publications/darpa99/index.htm 1998 BROADCAST NEWS BENCHMARK TEST RESULTS: ENGLISH AND NON-ENGLISH WORD ERROR RATE PERFORMANCE MEASURES, http://www.nist.gov/speech/publications/darpa99/pdf/ov10.pdf.

Pigeon, S. and Vandendorpe, L. "The M2VTS Multimodal Face Database (Release 1.00)", Proc. First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA '97), Crans-Montana, Switzerland, March 12-14 1997, pp. 403-409.

Program of Work, NATO C3 Board Communication Network Sub Committee Ad Hoc Working Group on Narrow Band Voice Coding.

Reynolds, D.A. “Comparison of background normalization methods for text-independent speaker verification,” in Proceedings of the European Conference on Speech Communication and Technology, pp. 963-967, September 1997.

Rohlicek, J.R. “Gisting Conversational Speech”, Proc. ICASSP 92, San Francisco, CA, Vol. 2, pp. 113-116, March 1992.

Schroeder, M.R. (1999). *Speech Processing*. in Signal Processing for Multimedia. J. S Byrnes (Ed.) IOS Press.

Shadle, C. and Damper R. (2001). *Prospects for Articulatory Synthesis: A Position Paper*. 4th ISCA Workshop on Speech Synthesis.

Spanias, A. “Speech coding: a tutorial review”, Proc. IEEE, Vol. 82, pp. 1341-1382, 1994.

Sproat, Richard (Editor) (1997). *Multilingual Text-to-Speech Synthesis – The Bell Labs Approach*. Kluwer.

Swail, C. and Kobierski, R. “Direct Voice Input for Control of an Avionics Management System”, Proc of the American Helicopter Society 53rd Annual Forum, Virginia Beach, Virginia, April 29 – May 1, 1997.

Swail, C. “Speech Recognition Tests of a Verbex Speech Commander in the NRC Bell 412 Helicopter”, to be published.

Teixeira, C., Franco, H., Shriberg, E., Precoda, K. and Sömnez, K. “Evaluation of Speaker’s Degree of Nativeness Using Text-Independent Prosodic Features”, Proc. of the Workshop on Multilingual Speech and Language Processing, MSLP5, Aalborg, Denmark, September 2001.

Teixeira, C., Franco, H., Shriberg, E., Precoda, K. and Sömnez, K. “Prosodic Features for Automatic Text-Independent Evaluation of Degree of Nativeness for Language Learners”, ICSLP 2000 – 6th International Conference on Spoken Language Processing, Beijing, China, October 2000.

Teixeira, C. “Reconhecimento de Fala de Oradores Estrangeiros”, PhD thesis, DEEC (IST-UTL), Lisboa, 1998.

Teixeira, C., Trancoso, I. and Serralheiro, A. “Recognition of Non-Native Accents”, Eurospeech 97, Proceedings, Rhodes, September 1997.

Teixeira, C., Trancoso, I. and Serralheiro, A. “Accent Identification”, ICSLP 96, The Fourth International Conference on Spoken Language Processing”, Philadelphia, October 1996.

Teixeira, C. and Trancoso, I. “Continuous and semi-continuous HMM for recognising non-native pronunciations”, Proceedings of 1993 IEEE Workshop on Automatic Speech Recognition, Snowbird, Utah – USA, December 1993.

Teixeira, C. and Trancoso, I. “Word rejection using multiple sink models”, Proceedings of International Conference on Spoken Language Processing, Banff, 2:1443-1446, 1992.

Test and Selection Plan 2400bps/1200bps Digital Voice Coder AC/322(SC/6-AHWG/3) NATO AdHoc Working Group on Narrowband Voice Coding Version – V1.14 – October 2000.

Tomokiyo, L.M. and Waibel, A. “Adaptation Methods for Non-Native Speech”, Proc. of the Workshop on Multilingual Speech and Language Processing, MSLP6, Aalborg, Denmark, September 2001.

Trancoso, I., Viana, C., Mascarenhas, I. and Teixeira, C. “On deriving rules for nativised pronunciation in navigation queries”, Eurospeech 99, Proceedings, Budapest, September 1999.

Trustech LTD, 9 October 2001, <http://www.truster.com/>

van Santen, J., Sproat, R., Olive, J. and Hirschberg, J. (Editors) (1997). *Progress in Speech Synthesis*. Springer.

Wong, E. and Sridharan, S. “Fusion of Output Scores on Language Identification System”, Proc. of the Workshop on Multilingual Speech and Language Processing, MSLP11, Aalborg, Denmark, September 2001.

XANDI Electronics, “Voice Stress Analysis Kit,” Instruction Manual, Model No. XVA250.

Zissman, M., Gleason, T., Rekart, D. and Losiewicz, B. “Automatic Dialect Identification of Extemporaneous conversational Latin American Spanish Speech”, Proc. ICASSP, Volume 2, pp. 777-780, Atlanta, 1996.

Zissman, M.A. “Predicting, diagnosing and improving language identification performance,” in Proceedings of the European Conference on Speech Communication and Technology, September 1997.

Zissman, M., van Buuren, R., Grieco, J., Reynolds, D., Steeneken, H. and Huggins, M. “Preliminary Speaker Recognition Experiments on the NATO N4 Corpus”, Proc. of the Workshop on Multilingual Speech and Language Processing, MSLP2, Aalborg, Denmark, September 2001.

Chapter 4 – ASSESSMENT AND EVALUATION

4.1 GENERAL PRINCIPLES OF SPECIFICATION AND ASSESSMENT OF SPEECH AND LANGUAGE PROCESSING SYSTEMS

4.1.1 Introduction

Very few speech and language processing applications involve “stand-alone” speech and language technology. Speech, handwriting and text provide essential components of the more general human computer interface alongside other input/output modalities such as pointing, imaging and graphics. This means that the actions and behaviours of the speech and language-specific components of a complex multi-modal “human-computer interface” (HCI) inevitably have to be orchestrated with respect to the other modalities and to the application itself, and this is usually achieved by some form of interactive dialogue process (simultaneously taking into account the wide range of human factors involved).

The complexity of the human-computer interface, and the subtle role of speech and language processing within it, has been (and continues to be) a prime source of difficulty in deploying speech and language systems in military applications. Not only are field conditions very different to laboratory conditions, but there has been a serious lack of agreed protocols for specifying such systems and for measuring their overall effectiveness.

This means that applications developers and system designers are unable to select appropriate “off-the-shelf” HCI components (such as automatic speech recognisers, for example) not simply due to a lack of standardised evaluation criteria for such system components, but also from a lack of a clear understanding of the implications of the performance of each system component on overall system effectiveness.

4.1.2 The Right Application using the Right Technology

One possible model for understanding the relationship between speech and language applications and the corresponding technology is illustrated in Figure 4.1. The key notion is that, not only is it necessary to match the “capabilities” of the technology with the “requirements” of the applications (and this can be done at either the technical or operational levels), but it is also important to emphasise that the purpose of introducing speech and language technology into an application is to achieve the appropriate operational benefits. The process illustrated in Figure 4.1 shows how, in order to specify and assess the relevant technology, the operational benefits being sought in current and future speech and language applications (such as manpower savings or increased mission effectiveness) have to be expressed either in terms of operational requirements (such as increased data entry rates or reduced head-down time) or in terms of technical requirements (such as >200 words-per-minute data entry rate or >95% word recognition accuracy). Likewise the technical features of current and future speech and language technology (such as hidden Markov modelling or neural networks) have to be expressed in terms of the corresponding technical or operational capabilities.

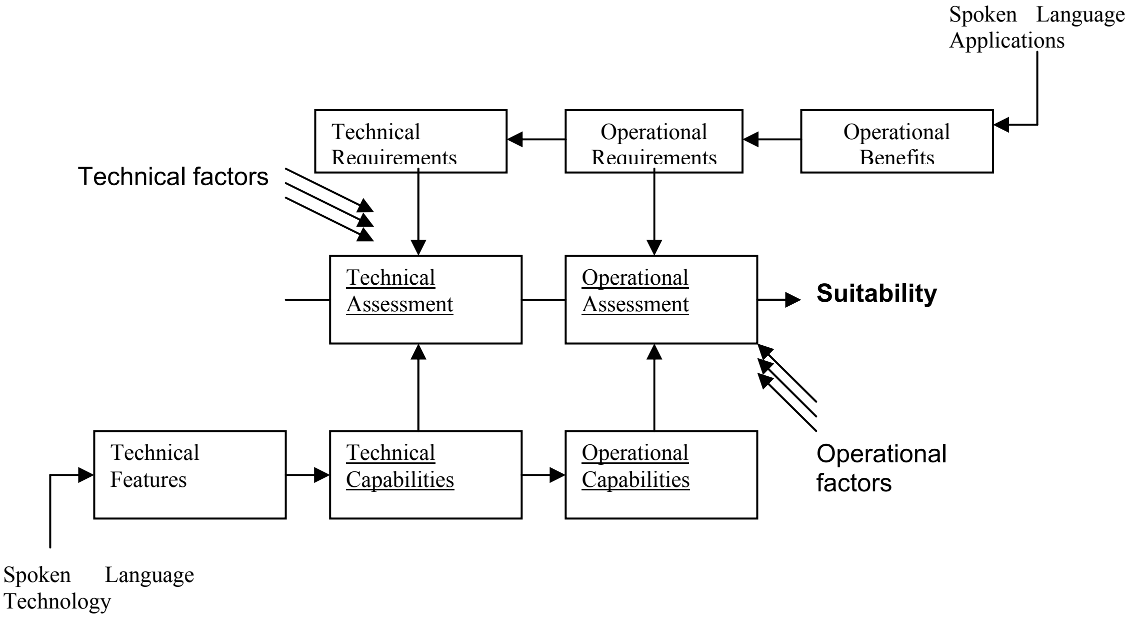


Figure 4.1: A Model of the Relationship between the Applications of Speech and Language Systems and the Underpinning Technology.

4.1.3 System Specification

Of course many different factors influence the suitability of particular speech and language systems for specific applications (such as the level of acoustic noise in the environment or the degree of training which the user has received), therefore the requirements and capabilities are expressed as multi-dimensional “profiles”; the “capability profiles” indicates the performance that is available (what can be done) and the “requirement profiles” indicates the performance that is required (what is needed).

As a consequence, the specification of a speech and language system inevitably involves a process whereby a system designer would have to provide qualified judgements about the required values and acceptable ranges (together with the weighting of each dimension in terms of its relative importance) against a comprehensive checklist of performance-related factors.

For example, for an automatic speech recognition system the list of technical requirements would have to specify the characteristics of all of the following influencing factors:

- Environment
- Talking Style
- Transducer
- Enrollment
- Channel
- Implementation
- Task
- Controls
- Speaker(s)

and also the following performance requirements:

- Recognition error rate
- Packaging
- Speed
- Weight
- Response time
- Power
- Enrollment time
- Cost
- Size

4.1.4 Evaluation and Assessment

The overall goodness of a speech and language system can be viewed as corresponding to a “performance envelope” in which performance in one dimension can be traded against performance in another. However, in many circumstances the most appropriate technology may not be available for a given application, and this leads to the concept of a best fit between requirements and capabilities (where some requirements may not be satisfied). Shortfalls in one dimension would have to be traded against gains in others, and adjustments to the capabilities (or indeed the requirements) would have to be made in order to fulfill some bottom-line criterion (such as minimum cost, for example).

In the assessment of speech and language systems it is possible to distinguish three main methodologies:

- Live “field” trials,
- Laboratory-based tests,
- System modelling paradigms.

The first of these of course is likely to provide the most representative results but, from a scientific point of view, there are likely to be a number of uncontrolled conditions and this limits the degree of generalisation that can be made from application to application. Field trials also tend to be rather costly operations to mount. Laboratory testing is per force more controlled and can be relatively inexpensive, but the main problem is that such tests may be unrepresentative of some (possibly unknown) key field conditions and give rise to the observed large difference between performance in the laboratory and performance in the field. The third possibility, which is itself still the subject of research, is to model the system (and its components) parametrically. In principle, this approach could provide for a controlled, representative and inexpensive methodology for assessment but, as yet, this area is not sufficiently well developed to be useful.

The term “assessment” also covers a range of different activities. For example, a suitable taxonomy of assessment activities should include: “calibration” (does the system perform as it should), “diagnosis” (how well does the system perform over a range of diagnostic conditions), “characterisation” (how well does the system perform under parametrically controlled conditions), “prediction” (how well will the system perform under different conditions) and “evaluation” (how well does the system perform overall). Of all these, the last evaluation has received a bulk of the attention in speech and language systems assessment.

It is also the case that assessment protocols are required which address a large number of different types of speech and language systems. For example, such systems range from laboratory prototypes to commercial off-the-shelf products, from on-line to off-line systems, from standalone to embedded systems, from sub-systems to whole systems and from speech and language systems to speech and language based HCI systems.

The majority of activity in the area of speech and language system assessment has concentrated on evaluating system components (such as measuring the word recognition accuracy for an automatic speech recogniser, for example) rather than overall (operational) effectiveness measures of complete HCI systems. Since the publication of the US National Bureau of Standards guidelines in 1985, there have been considerable developments at the international level. In Europe, the EU Speech Assessment Methods project (SAM) established a standard test harness for both recognisers and synthesizers and in the US a very effective assessment paradigm has been funded by DARPA which included an efficient production line of “hub and spoke”-style experiments involving the co-ordinated design, production and verification of data, distribution through the Linguistic Data Consortium, and with the National Institute for Science and Technology responsible for the design and administration of tests and the collation and analysis of the results.

These activities emphasis the importance of “bench marking”, either through the implementation of standard tests, or by reference to human performance or to reference algorithms.

4.1.5 Standards and Resources

The requirement for agreed standards and guidelines pervades all of the links in the speech and language system R&D chain starting from the research community (for algorithm development and benchmarking), to product developers (for performance optimisation), system integrators (for component selection), manufacturers (for quality assurance), sales staff (for marketing), customers (for product selection) and users (for service selection).

In addition, many of the speech and language technologies rely heavily on the availability of substantial quantities of speech and language corpora: first, as a source of material from which to derive the parameters of the constituent models, and second, in order to assess performance under controlled (repeatable) test conditions.

The most significant activity on speech and language standards and resources has been the SAM project which ran from 1987 to 1993. The SAM project arose out of the need to develop a common methodology and standards for the assessment of speech technology systems which could be applied within the framework of the different European languages.

4.1.6 Corpora

Three types of speech and language corpora are typically of interest: “analytic-diagnostic” material which is of primary importance to progress in basic science and which is specifically designed to illuminate specific phonetic and linguistic behaviour, “general purpose” material which includes vocabularies which are either common or which are typical of a wide range of applications (for example, alpha-numeric words or standard control terms), and “task-specific” material which reflects different levels of formalised spoken monologue/dialogue within constrained discourse domains.

Clearly general purpose corpora are easy to collect and are useful in a general sense but, of course, they have only limited practical value. On the other hand, although task-specific corpora can be time-consuming to collect and are only relevant to a specific domain, they are obviously directly useful for the purposes of practical applications. Diagnostic corpora are time consuming to design, but they are extremely useful for research purposes.

The availability of standard corpora is of great importance for the speech community and a number of national and international bodies have been responsible for co-ordination, distribution and production of appropriate databases. For military applications, NATO Research Study Group on Speech and Language Technology (AC323/IST/RTG-001, formerly RSG.10) has, since the late 1970s, provided an effective mechanism for exchanging information on spoken language standards and resources between Canada, France, Germany, the Netherlands, the UK and the USA. RSG.10 was responsible for the first publicly available multi-lingual speech corpus, and has subsequently released on CD-ROM a database of noises from a range of selected military and civil environments (NOISE-ROM) and related experimental test data (NOISEX). In addition RTG-001 maintains a database of information about speech corpora specifically related to the military needs. At time of writing the database contains about 40 entries, some of which include non-speech sounds such as background noise in military vehicles.

4.1.7 Other Resources

The Linguistic Data Consortium (LDC) was founded in the US in 1992 to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. The consortium distributes previously-created databases, as well as funding and coordinating the funding of new ones. The LDC is closely tied to the evolving needs of the community it supports and has helped researchers in several countries to publish and distribute databases that would not otherwise have been released.

In Europe, the European Language Resources Association (ELRA) was established as a non-profit organization in 1995, with the goal of creating an organization to promote the creation, verification, and distribution of language resources in Europe. Eventually, ELRA will serve as the European repository for EU-funded language resources and interact with similar bodies in other parts of the world (such as the LDC).

Another valuable resource is a series of handbooks of speech and language standards and resources produced by the Expert Advisory Group on Language Engineering Standards, sponsored by the EU. The initiative covers a wide range of topics including methodologies for the creation and interchange of electronic language resources such as text and speech corpora, computational lexicons and grammars formalisms, and the evaluation and quality assessment of language processing systems and components (see Gibbon et al., 1997 and Gibbon et al., 2000).

4.2 SPECIFICATION AND ASSESSMENT OF SPEECH COMMUNICATION SYSTEMS

4.2.1 Introduction

Evaluation of speech related aspects of communication systems can be separated in intelligibility, response in adverse conditions (jamming, channel noise, background noise, etc.) and the response time (delay). Generally the intelligibility or speech quality is measured under various representative conditions. Criteria for the performance under optimal conditions (back-to-back connections) and representative usage conditions are proposed. Measures of performance and intelligibility can be determined both by subjective and objective methods. There are three groups of measuring methods:

- Subjective intelligibility measures based on phonemes, words or sentences;
- Subjective quality measures related to a global impression; and
- Objective measures based on physical aspects of the speech signal or the speech transmission system.

Subjective intelligibility measures are in general very representative as speakers and listeners are used. However, to obtain reproducible results, much effort is required to perform the measurements and results are dependent on the speech material used for the test. Also no diagnostic information is obtained. One simple method is based on quality rating where listeners score their impression of the speech quality. This is a global method and requires many listeners. A recent development is presented with the speech communicability test which evaluates the performance of an actual channel by using real conversations.

Finally objective methods were developed in which the transmission quality is derived from physical parameters. These methods are easier to apply and offer in addition to the prediction of intelligibility also useful diagnostic information. Unfortunately these methods cannot be used for voice coders like LPC-based systems.

The "Ergonomic Assessment of Speech Communications" will be internationally standardised by a new ISO standard (ISO 9921).

4.2.2 Intelligibility Measures and Quality Rating

A number of subjective intelligibility tests have been developed for the evaluation of speech communication channels. In general, the choice of the test is related to the purpose of the study: are systems to be compared or rank-ordered, are systems to be evaluated for a specific application or must the development of a system be supported. For both types of application a different test may be appropriate. An overview focused on the assessment of speech processing systems is given by Steeneken (1992).

Subjective intelligibility tests can be largely categorised by the speech items tested and by the response procedure used. The smallest items tested are at the segmental level, i.e. phonemes. Other test items are CV, VC, and CVC combinations (C=consonant, V=vowel), nonsense words, meaningful words, and sentences.

Besides intelligibility scores, speech quality can also be determined by questionnaires or scaling methods, using one or more subjective scales such as: overall impression, naturalness, noisiness, clarity, etc. Speech quality assessment is normally used for communications with a high intelligibility, since most tests based on intelligibility scores are inappropriate because of ceiling effects.

4.2.2.1 Tests at Phoneme and Word Level

A commonly used test for determining phoneme scores is the rhyme test. A rhyme test is a forced-choice test in which a listener, after each word that is presented, has to select his response from a small group of visually presented alternatives. In general, the alternatives only differ with respect to the phoneme at one particular position in the test word. For example, for the Dutch language and for a test with a plosive in the initial consonant position, the possible alternatives might be: Bam, Dam, Gam, Pam, Tam, Kam. A rhyme test is easy to apply and does not require much training of the listeners. Frequently used rhyme tests are the Modified Rhyme Test (MRT, testing consonants and vowels) and the Diagnostic Rhyme Test (DRT, testing specific initial consonant pairs only).

A more general approach is obtained with a test with an open response, such as with monosyllabic word tests. Open response tests make use of short nonsense or meaningful words most often of the CVC type.

The test results can be presented as phoneme scores and word scores but also as confusions between the initial consonants, vowels, and final consonants.

The confusion matrices obtained with open response tests provide useful (diagnostic) information for improving the performance of a system.

4.2.2.2 Tests at Sentence Level

Sentence intelligibility is sometimes measured by asking the subjects to estimate the percentage of words correctly heard on a 0 to 100% scale. This scoring method tends to give a wide spread among listeners. Sentence intelligibility saturates to 100% at poor signal-to-noise ratios, the effective range is small (see Figure 4.2).

4.2.2.3 Quality Rating

Quality rating is a more general method, used to evaluate the user's acceptance of a transmission channel or speech output system. For quality ratings, normal test sentences or a free conversation are used to obtain the listener's impression. The listener is asked to rate his impression on a subjective scale such as the five-point scale: bad, poor, fair, good, and excellent. Different types of scales are used, including: intelligibility, quality, acceptability, naturalness etc. Quality rating or the so-called Mean Opinion Score (MOS) gives a wide variation among listener scores. The MOS does not give an absolute measure since the scales used by the listeners are not calibrated. Therefore the MOS can be used only for rank-ordering conditions. For a more absolute evaluation, the use of reference conditions is required as a control.

4.2.2.4 Objective Intelligibility Measures

The first description of the use of a computational method for the prediction of the intelligibility of speech and its realization in an objective measuring device, was developed in 1959 by Licklider. Presently a measure based on the Speech Transmission Index (STI, Steeneken and Houtgast 1980, 1999) is standardised by IEC 60268-16 (1999).

The method assumes that the intelligibility of a transmitted speech signal is related to the preservation of the original spectral differences between the speech sounds. These spectral differences may be reduced by band pass limiting, masking by noise, non-linear distortion components, and distortion in the time domain (echoes, reverberation). The reduction of these spectral differences can be quantified by the effective signal-to-noise ratio, obtained for a number of relevant frequency bands. As the STI is focused on the reproducibility of the spectral and temporal envelope and does not take into account the reproducibility of the carrier, the method cannot be applied to vocoders.

For the application of the STI method a specific test signal is applied at the input side of the system under test. An analysis is made of the output in order to obtain the effective signal-to-noise ratios for all frequency bands considered (seven octave-bands ranging from 125 Hz to 8 kHz). The test signal and the analysis is designed in such a way that non-linear distortion and distortion in the time domain affect the information content of the test signal in a manner similar to the degradation of speech. A weighted contribution of the measured information transfer in the seven octave bands results in a single index, the STI. The measuring method and the algorithm have been optimized for an optimal correlation of the STI with the subjective intelligibility.

4.2.2.5 Criteria and Relation between Various Measures

Figure 4.2 shows, for five subjective intelligibility measures, a quality rating and the relationship with the objective STI. Also shown are comparable signal-to-noise ratios. This illustrates the effective range of each test. The given relation between intelligibility scores and the signal-to-noise ratio is valid only for noise with a frequency spectrum similar to the long-term speech spectrum. In this instance a voice-babble is used. A signal-to-noise ratio of 0 dB then means that speech and noise have an equal spectral density.

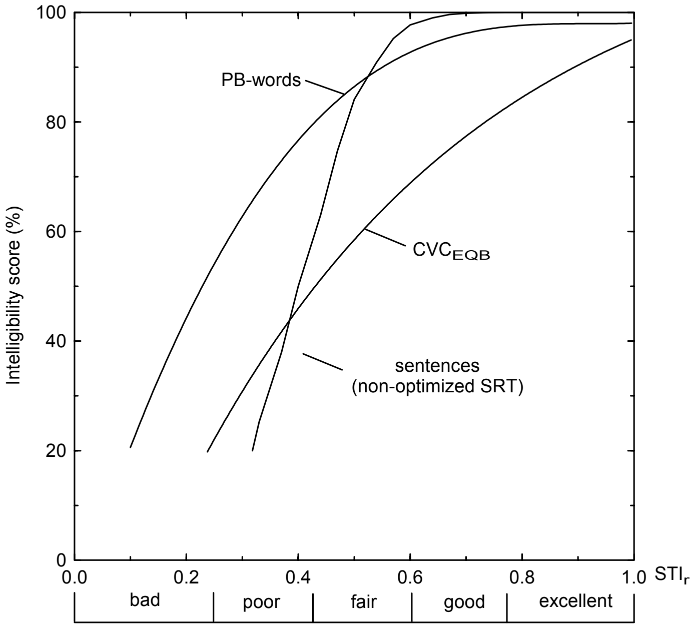


Figure 4.2: Relationship between some Intelligibility Measures according to Steeneken (1992) and DIS ISO 9921. (Noise with a spectrum shaped according the long-term speech spectrum.)

As can be seen from the figure, the CVC-nonsense words discriminate over a wide range, while meaningful test words have a slightly smaller range. The digits and the alphabet give a saturation at a signal-to-noise ratio of 5 dB. This is due to: (a) the limited number of test words and (b) the fact that recognition of these words is controlled mainly by the vowels rather than by the consonants.

In general for military communications a back-to-back performance qualified as good is required. This corresponds with a minimum CVC-word score of 70%, a DRT of 96% or an STI of 0.6. The criterion for just acceptable in the worst condition and qualified as poor (e.g. related to a condition that less than 100% intelligibility of digits and redundant sentences is obtained) corresponds to a CVC-word score of ~50% or an STI of 0.45.

4.2.3 Speech Communicability Testing

Quite often, the performance of speech communication channels (or systems) may be evaluated adequately by only investigating speech quality or speech intelligibility. This is not always the case; for example, long transmission delay times will make communication clearly more difficult, even if the intelligibility of the transmitted speech is not affected.

Speech communication performance tests that include such effects are known as *speech communicability tests*. TNO has developed a speech communicability test that measures two separate performance indicators: the *efficiency* of communication, and the *acceptability* of the communication channel. The test is based on real-time communication between test subjects. Pairs of subjects communicate using the communication channel under test, and are given a joint task that forces them to communicate. This task is basically a card game, derived from the popular gambling game ‘Black Jack’. The structure of the conversation between the test subjects (‘players’) is quite predictable, due to the rules of the game. This makes it possible to assess the efficiency of their communication, by measuring the time they need to complete certain parts of the game. The acceptability of the communication channel is assessed through post-hoc questionnaires.

An important variable that influences the efficiency of communication, is the influence of context: simply communicating single digits (which lets the listening party choose from a set of only 10 options) is by far easier than communicating completely unpredictable phrases. The communicability test uses certain ‘key words’ (or phrases), which the subjects need to communicate about their cards in the card game that they are playing. By changing the sets of key words, the degree of redundancy of the communicated speech is manipulated.

An important feature of the test is the fact that it allows for compensating strategies. Under adverse conditions, such as high levels of background noise, people automatically adjust their speech to compensate for this (raise their vocal effort, for instance). The effect of such strategies is included in the results of the communicability test, whereas it is not included when using straightforward intelligibility or quality tests. This is illustrated by figure 4.3.

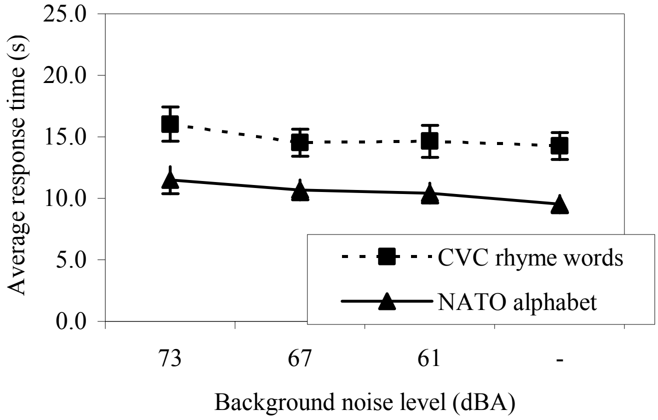


Figure 4.3: Average Subject Responses Time on Part of the Task (representative for communication efficiency) for Four Different Levels of Background Noise (average speech level at normal vocal effort was 70 dB(A)). Scores are based on 8 subject pairs, 50 observations per condition per pair. The error bars represent standard errors.

Figure 4.3 was obtained by conducting communicability tests with two different kinds of key-words: the NATO spelling alphabet (highly redundant) and CVC rhyme words (nonsense words, virtually without any redundancy). There is a clear and significant difference between the scores for both word categories. Figure 4.3 shows the absence of an effect on efficiency when changing the level of background noise, where a straightforward intelligibility would have predicted such an effect. Figure 4.3 is representative of what happens in practice: the test subjects have dealt with the background noise by adapting their speech production. This is likely to affect the acceptability of the channel; this is reflected by the results presented in figure 4.4.

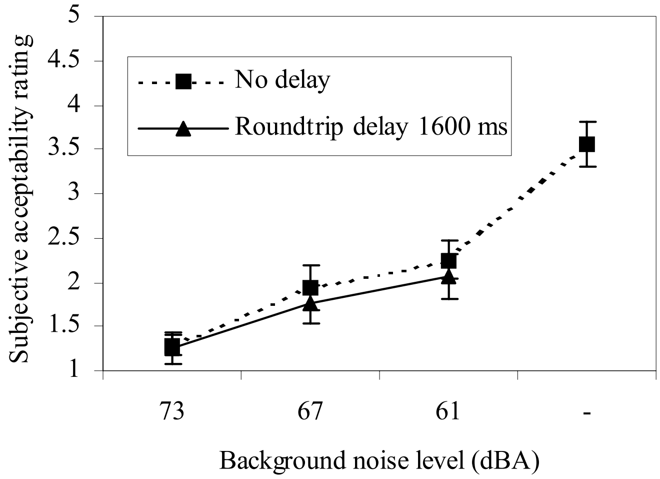


Figure 4.4: Average Acceptability Rating for Four Different Levels of Background Noise (average speech level at normal vocal effort was 70 dB(A)). Results are based on responses from 16 subjects. The error bars represent standard errors.

Figure 4.4 shows a significant effect of background noise on acceptability, both in absence and in presence of transmission delay. By combining results from figures 4.3, 4.4, we know what to expect in practice.

Figure 4.5 demonstrates that transmission delay does have an effect on communication efficiency. The results shown in this figure are similar to those of figure 4.3, only this time for different values of the roundtrip delay time instead of the background noise level.

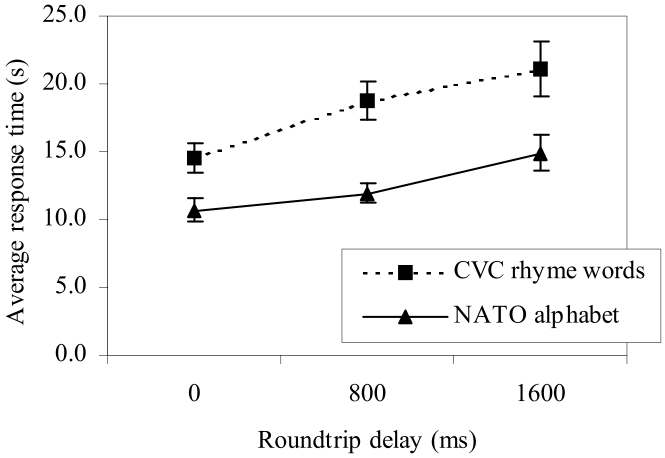


Figure 4.5: Average Subject Responses Time on Part of the Task (representative for communication efficiency) for Three Different Roundtrip Delay Values. Scores are based on 8 subject pairs, 50 observations per condition per pair. The error bars represent standard errors.

The effect of roundtrip delay on communication efficiency as shown in figure 4.5 is statistically significant. Again, the redundant words (NATO alphabet) show a higher efficiency than the (nonsense) CVC rhyme words.

We may conclude that the communicability test is a useful tool for predicting speech communication performance in practice, especially for communication channels featuring delay.

4.3 SPECIFICATION AND ASSESSMENT OF SPEECH RECOGNITION SYSTEMS

4.3.1 Introduction

In this context, the term "recognition systems" includes systems designed to recognise words, speakers, accents or languages. Most such systems have a common technological basis, but use different training methods or detection criteria to distinguish the different target attributes of the speech with which they are presented. Each of these different systems will also have specific requirements for evaluation, and in general these will be application dependent. In the present document, it is possible to describe only the broad features of such evaluations. For a detailed discussion, the reader is directed to Volume 3 of the "EAGLES" handbook, "Spoken Language System Assessment".

An assessment of a recognition system is necessarily a *sample* of its performance, and care must therefore be taken to ensure that the sample is representative of the application, or, at least, randomized so as to minimise the possibility of bias. A further requirement is that the sample is large enough to limit the variance in the results to an acceptable level. For a small vocabulary word recognition task with a limited and known user population, it is just possible that a test could include all possible words and syntax paths spoken by all the users, though this will seldom be the case. Even then, the test will not include all the possible *utterances* that the recogniser may meet. In the usual case, it will be necessary to select a sample of speakers and utterances, and the specification and testing should take account of the statistical uncertainties arising from sampling.

In the case of a speaker-dependent recogniser, it should be remembered that the data used to train the speech models are also a sample of possible training data. A further consideration arises if the system is likely to be subject to repeated testing, during development for example. If the same test material is used repeatedly, there is a danger that the system will become "tuned" to deal with peculiarities in the data, and may therefore perform worse in the real application. If repeated testing is necessary, the test corpus should be divided into sections to be used at each stage, or at least to keep one set of data completely unused until the final acceptance test.

4.3.2 Recognition Systems

4.3.2.1 Word Recognition

The technical assessment of a speech recogniser will usually be carried out offline, using pre-recorded speech. It will generally consist of recording the responses to input of words or sentences which are structured according to the task for which the recogniser is required, but other more general material may also be used.

It is easy to focus on word recognition accuracy as the main figure of merit, but it should be borne in mind that other considerations will play a part in the overall operational suitability of the equipment.

Although high accuracy is obviously desirable, it is only one of the *technical capabilities* of the recogniser, and may not be a dominant factor in the overall benefits of the system. The needs of any given application must be determined in the context of the whole interface.

The main requirements to be considered fall into three broad classes. Firstly, there are the requirements of the task – vocabulary, syntax, speaking style, etc. The size of the vocabulary is the main determinant of many of the characteristics of the recogniser and of methods of assessment. Secondly, there are factors connected with the user population – mainly arising from whether the recogniser is required to be speaker-dependent or speaker-independent. Thirdly, there are considerations of signal quality – microphone and channel characteristics, noise, and environmental factors like vibration or G-force which affect the user in producing the speech.

Performance measures used in the technical assessment may include word accuracy, phrase accuracy, response time, and robustness. Given a suitable assessment database, word and phrase accuracy are relatively easy to determine (within the precision allowed by statistical considerations). Depending on the application, it may also be necessary to determine the response to out-of-vocabulary utterances. Response time can be difficult to define, since many speech sounds do not end abruptly but die away over a period of tens or even hundreds of milliseconds, and its measurement requires a database that is labelled at a suitable level, which may need considerable effort to produce.

"Robustness" is a term used to describe the response of a system when its inputs are degraded in some way. A system which performs very well under ideal conditions, but poorly under other conditions is unlikely to be useful in military applications, in which noise, stress, and poor channel characteristics are usual. The direct performance measures, such as accuracy and response time, should therefore be measured under a range of conditions suited to the application and the results should be assessed in terms of their likely operational effects.

The operational assessment will generally require a simulation of the workstation which is the target application, if not a proper field trial. A speech recogniser will seldom, if ever, be the only component of a human-machine interface, so the evaluation may need to take account of the other input and output modalities and their interactions with each other and the task to be carried out. The human subjects used in the tests will need to be trained to use the speech recogniser, and in the overall task, for a fair assessment.

Functional measures used in the operational assessment will vary depending on the application, but will often include transaction time and accuracy, user satisfaction, and user workload. In many cases, the benefits of speech recognition will result in improved performance of a primary task, which may not be directly related to the control inputs carried out using speech. An example is found in military cockpits, where the primary task for a pilot is *always* "to fly the aircraft", but other control actions are often necessary or desirable. The effectiveness of a speech recogniser could be measured in terms of a reduction in the amount of time that the pilot needs to look at his instruments inside the cockpit or take his hands off the flying controls. Even if the recognition accuracy seemed to be poor, it could still allow the pilot to change instrument settings without affecting his ability to monitor and control the aircraft's flight path, and would therefore be beneficial at the operational level. At this level, each application will have its own requirements for assessment, so it is not possible to give guidelines here.

4.3.2.2 Speaker Classification

"Speaker classification" is a general term covering speaker identification and speaker verification, and can also be applied to language and accent identification and a variety of less common applications such as gender or age identification. Speaker verification systems will usually be used to control access to a secure area or system, so the specification and assessment must take particular account of the consequences of failure. The false rejection of a genuine user may be no more than a nuisance, but false acceptance of an impostor could have very serious repercussions. On the other hand, if a speaker identification system is to be used in a surveillance operation to select radio transmissions for further study by human operators, a high level of false acceptances may be required in order to minimise rejection of genuine utterances of the target speaker.

The design of a database for evaluating such systems requires attention to some special features, particularly to the avoidance of unrelated features which may be correlated with speaker identity by chance. Recordings made in relatively uncontrolled circumstances, such as over the telephone network, are especially prone to these problems. If each speaker in the database is always recorded from the same telephone handset but different speakers use different handsets, then the system may be identifying the handset rather than the speaker. A similar consideration applies to any transmission channel or even to background noise present during the recording.

Where a system is required to operate over an extended period, it must take account of changes in an individual's voice characteristics which take place on a time scale of hours, days or months. Normally, the speech used to register a new user with the system will be collected in one session, and the speech to be evaluated will be produced at intervals over an extended period of time following the registration. The design of an evaluation database should follow this general pattern. If the training and test data are all recorded in one session, it is probable that the results from the test will not be representative of the performance of the system in real operations.

Measurement of the performance of a speaker classification system is more complicated than that for recognition systems, because they always have thresholds that can be adjusted to change the balance between false acceptances and false rejections. Full characterisation of a system would require measurement of its performance over the whole range of threshold settings, or at least over a range likely to be of interest in practice. This leads to the use of the *Receiver Operating Characteristic*, (the term being derived from Communication Theory) where the rate of false acceptance is plotted against the rate of false rejections, with the threshold as the parameter which generates both values. Figure 4.6 shows an example of an ROC, which has an approximately hyperbolic form. If a single figure-of-merit is required, the Equal Error Rate (EER) figure is usually quoted. This represents the case when the rate of false acceptances is equal to the rate of false rejections, and is found where the ROC curve intersects the diagonal of unit slope, as shown in the Figure. In this case the EER is 20%. The EER is, of course, a great simplification of a complex characteristic, but it does provide a concise measure of performance in the vicinity of the equal error rate. Performance at points where the error rates are very different, as may be required in an access control system, may not be closely related to EER.

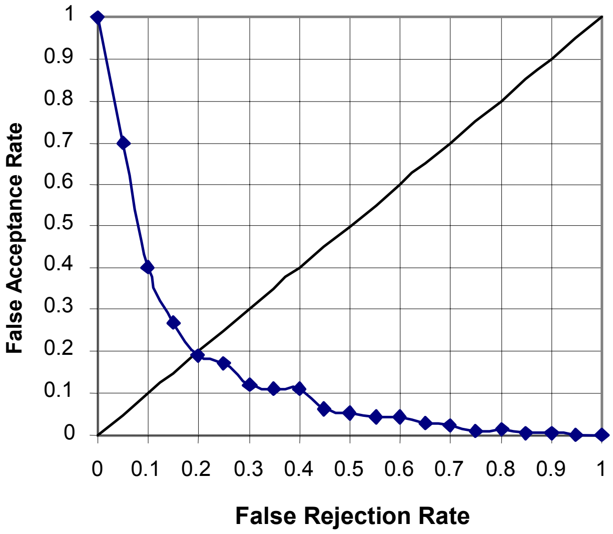


Figure 4.6: Receiver Operating Characteristic for Speaker Identification System.

4.4 REFERENCES

ISO 9921, *Ergonomics – Assessment of speech communication (Draft International Standard, 2001).*

Licklider, J.C.R., Bisberg, A. and Schwartzlander, H. 1959. "An electronic device to measure the intelligibility of speech", Proc. Natl. Electronic Conf. 15, pp. 329-334.

IEC 60268 - 16 (1998). *Sound system equipment – Part 16: The objective rating of speech intelligibility by speech transmission indexcommunicability tests*.

Steeneken, H.J.M. and Houtgast, T. (1980). "A physical method for measuring speech-transmission quality", J. Acoust. Soc. Am., **67** (1), pp. 318-326.

Steeneken, H.J.M. (1992). "Quality evaluation if speech processing systems", Chapter 5 in *Digital Speech Coding: Speech coding, Synthesis and Recognition*, edited by Nejat Ince, (Kluwer Norwell USA), pp. 127-160.

Steeneken, H.J.M. and Houtgast, T. (1999). "Mutual dependence of the octave-band weights in predicting speech intelligibility". Speech Communication 28, pp. 109-123.

Gibbon, D., Moore, R. and Winski, R. Editors (1997). *Handbook of Standards and Resources for Spoken Language Systems,* "Assessment of recognition systems", pp. 381-407. Mouton de Gruyter, Berlin.

Gibbon, D., Mertens, I., and Moore, R. Editors (2000). *Handbook of Multi-model and spoken dialogue systems,* "Consumer off-the-shelf (COTS) speech technology product and service evaluation", pp. 204-239. Kluwer Academic Publisher, Berlin. (ISBN 0-7923-7904-7).

Chapter 5 – CASE STUDIES AND (FUTURE) APPLICATIONS

In this chapter, applications of speech and language technology that are already in service or likely to be in the near future are presented. The selection of contributions is not intended to be exhaustive, but shows the level of maturity of the main technology areas and covers a wide spread of applications. Automatic speech recognition frequently draws the main attention, but it should not be forgotten that coding, synthesis and other technologies also have their applications, and indeed are more advanced in terms of actual use. Secure speech transmission using LPC-10 at 2400 Bps has been in use for many years and saw active service in the Gulf War. Similarly, voice warnings are in use in many aircraft cockpits, both military and civil.

Speech recognition has perhaps the widest range of possible applications. Its use is almost imperative in modern single-seat fighter and strike aircraft, where system complexity makes the workload very high. It finds broader use as part of a multi-modal interface in operations room applications, which might be aboard ship, on land or in the air. The object is to reduce the operator's workload or to enable him to handle more data. In training situations, such as for air-traffic controllers, the main motivation is the saving of the time of highly trained personnel who would otherwise have to play the parts of the aircraft being controlled.

In the communications area, reductions in the data rate required for speech transmission will aid covert operations, or allow secure, long-range communication with mobile units over HF radio. Maintenance of the quality and intelligibility of the speech is vital, as are suitable means of measuring these factors. The applications described below are mostly demonstrations or experimental systems; in several cases, however, it appears that only the will and the money would be required to put the systems into service. The technology now has adequate performance for useful application in many areas.

5.1 COCKPIT FAST JET

The ever-increasing complexity of aircraft systems coupled with requirements to operate at very low level and in all weathers creates a high workload in military cockpits, especially in single seat aircraft. A pilot's top priority should be to fly the aircraft, which requires the use of his hands and eyes. The operation of other equipment, although necessary for the mission, may be a distraction from the primary task. It has long been recognized that automatic speech recognition could provide a means of alleviating the workload and increasing eyes-out and hands-on time. Research suggests that this could have a significant impact on safety and mission effectiveness.

The military cockpit is however a very difficult environment for a speech recognizer. There is a high level of background noise and many factors that cause variations in the pilot's voice. On the other hand, there is a requirement for a very high level of accuracy: the pilot must be confident that the aircraft systems will respond, as he desires. This has created an impression in some quarters that the technology will never be good enough, but progress continues to be made. For example, special algorithms can be used for recognition in high levels of noise, achieving near 100% accuracy at 0 dB speech-to-noise ratio. Other developments are making recognizers more robust to variations in speech.

Successful implementation of automatic speech recognition in fast-jet cockpits will require careful attention to the human factors aspects. Not all tasks in the cockpit are suited to voice input. Once the appropriate tasks have been chosen, the vocabulary and syntax must be designed, and suitable feedback methods implemented. Above all, voice input must be regarded as an integral part of the cockpit design, and not as an optional extra. Recent results from trials in several countries indicate that the technology can deliver adequate performance for the more consistent speakers in less demanding flight conditions. There are reasonable grounds for expecting that the performance envelope can be expanded sufficiently to realize the benefits outlined above.

For example, research projects have been conducted on Automatic Speech Recognition applications in state-of-the-art single seat fighter cockpit in several countries. Cockpit tasks that can be executed with ASR include enhanced data entry operation for the Communication, Navigation and Identification (CNI) systems, and display management and control functions. Furthermore, the ASR applications include tasks supporting the normal Hands-On-Throttle-And-Stick (HOTAS) concept and interactive "crew assistant" applications using voice feedback. The primary pilot feedback mechanism is a normal avionics system response. A Press-To-Talk (PTT) switch is used to address the ASR system. The syntax and vocabulary are typically over 250 vocabulary words.

5.2 HELICOPTER

Military helicopters, as well as other military aircraft, become more and more complex because of the increased capability to perform a complex mission. Recent extensions have included more sophisticated sensors and diversified weaponry. A modern helicopter is a part of a complex system in which command posts, armored vehicles, other aircraft, and real-time intelligence centers work together.

The complexity of the systems embedded into the aircraft implies development of new concepts of the Man-Machine Interface. The tactical piloting task, under adverse conditions (stress, night, bad weather, vibration, and noise), gives a high workload for the crew. For this reason the interfacing to systems within the cockpit has to be simple. This also will prevent operating errors.

In this framework, at Sextant Avionique of France, a speech recognizer is applied in a tactical management system. The system includes an on-board station and a ground station. Both stations share a common mapping of the local area. This implies the management of a cartographic database and of additional objects that are related to the military units in the area.

By using a specific speech recognition system (Topvoice), the man-machine interfacing of the embedded system was drastically simplified. All operations dealing with visualization and configuration (zooming, zone-to-be-displayed, map layers display), and the operations for tactical object management (add, move, delete, tagging) are made by direct voice input.

Numerous tests during in-flight conditions have been completed in 1995 by more than twenty different speakers. The performance of the voice input system was 95% correct responses to the spoken commands even under adverse environmental conditions (windows open and maximum engine power).

The application has proven that the concept of voice input for military helicopters can be extended to others systems that require discrete operating control:

- System interface such as inquiries to get status from sub-systems of the aircraft (engine status, fuel management);
- Sensor suite management;
- Non-decisive actions concerning the weaponry.

Military aircrews are not the only ones who have complex and demanding tasks to carry out. The police use helicopters for a number of tasks, such as following vehicles containing suspects or searching for lost children, that demand close attention to television or infra-red images and map displays, while maintaining contact with as many as six other units via the radio. Should the observer take his eyes off the screen for a few seconds to change the radio channel or map scale, he could lose track of the suspect or miss a hot spot in undergrowth that could be the lost child. This task would often be undertaken by two crew members, in addition to the pilot.

In 1997, the Air Support Unit of the Devon and Cornwall Constabulary in the UK purchased a new helicopter and funded a feasibility study into the use of speech recognition to reduce the observer's workload. The study was carried out by DERA Farnborough, and included making recordings of the noise and vibration in the helicopter. These recordings were then used to reproduce the environment of the aircraft's cabin in DERA's man-carrying three-axis vibration rig when training the recogniser. In consultation with police observers, a vocabulary and syntax were developed for control of the radios, map displays and TV/IR display. In addition to the control words (numbering about 40), digits, and letters (using the ICAO phonetic alphabet), about 30 words such as "street", "road", "avenue", etc., and a similar number of town names were included. The names of smaller towns and villages could be spelt using the letters. The total vocabulary size was about 160 words. Advice was also given to the company that was installing the electronic equipment on the aircraft on the requirements for audio and other interfacing to the recogniser. The initial installation used a commercial continuous-speech, whole-word recogniser loaned by DERA, which had not been modified in any way to suit the application.

Three police observers (two male, one female) spent one day each at DERA Farnborough, recording speech for training the recogniser, and being trained in the use of the recogniser. Careful attention to reproducing the noise and vibration conditions of the real aircraft during training, and to building good models for the recogniser, resulted in a recognition accuracy of over 98%. These officers subsequently used the system in the course of their work, and considered it a great benefit, as it allowed them to concentrate their visual attention on the images and maps while still being able to change radio channels and operate other equipment. There was no requirement for a third crew member with this installation, which is believed to be the first operational use of airborne speech recognition by either military or police anywhere in the world.

5.3 SONAR

In the beginning sonar systems were very simple as described for example by Leonardo da Vinci: "If you place the head of a long tube in the water and place the other extremity to your ear, you will hear ships at a great distance from you." Since that time, the science of underwater acoustics has been refined and is still progressing rapidly. Today more and more often the complexity level and the amount of information available at a sonar suite output are exceeding the capacity limits of the classical man-machine interface. The main issues are:

- The greater detection range capabilities (sometimes more than the radar range) and consequently the increased number of detected contacts;
- The increased amount of information extracted for each contact (position, broad band and narrow band spectral description, transients, intercepted active sonar pulses);
- The greater algorithm complexity and consequently the increasing number of parameters; and
- The increasing number of graphical interactive tools aiming to help sonar operators in their various tasks. This includes contact tracking, data fusion, classification, contact motion analysis (in particular in the bearing only passive mode), situation and threat assessment (for a ship or for a zone), decision about maneuvers, weapons and sonar suite use.

It is worth noting that although high-level information is automatically provided by the sonar system, the interpretation by sonar analysts is still frequently required, for example analysis of low-level spatio-temporal signals. Finally in the ASW domain (anti-submarine warfare) everything is slower than in the air defense domain but paradoxically that does not really simplify the operators task because they must take into account all the various information gathered during a rather long period (a few hours).

Currently on board submarines, surface ships or maritime patrol aircraft, the sonar operators are overloaded. Because of the high training level and skills required for analysts, it is costly to increase the number of operators. Two solutions may be feasible to improve the performance of the operators by improving the various sonar analysis algorithms and by improving the efficiency of the man-machine interaction. This last point constitutes an opportunity for using man-machine communication by voice, integrated to the use of the standard display, keyboard and trackball (or joystick).

At present, voice input is not used operationally in the ASW domain. However, experimental studies have been carried out to integrate the speech recognizer into the more general man-machine interaction system. Voice input can be integrated in the following tasks:

- Panoramic surveillance on board submarines or ASW surface ships. This includes management of the intermittent tracks and contacts provided by the sonar suite, association/fusion of tracks, and recording of information about each contact (such as acoustic data, behavior, classification, crossing with other tracks). Speech recognition can also be useful for controlling the interactive display. For example requesting the intermediate results from raw signals, detected events, lofar, etc.
- Classification tasks performed on board ASW ships or aircraft or in shore-based intelligence centers. Tools for the analysis of the contact signature and the matching to known submarine or surface ship signatures are to be controlled by voice. Voice input can also be useful for controlling the display itself.
- Control of the various interactive graphical tools used for situation and threat assessment, and the decision aids concerning maneuvers, weapons and sensors.
- ASW tactical training systems to be used in land-based training centers.

In all these sonar related tasks, voice input based on connected word recognition provides the following benefits:

- The user can look continuously at the object without looking at menus or the keyboard.
- The cursor remains available for pointing out objects on the screen rather than pointing out menu items.
- The screen surface allocated to menus is reduced.
- Consequently the interaction can be faster.

These advantages become more important as the Control Information and Command (CIC) room has generally a low light level and as the use of the cursor (controlled by the trackball) and of the keyboard can be difficult in this situation especially when the ship is rolling and pitching.

It is worth noting that the CIC layout will probably be slightly modified by the introduction of man-machine communication by voice: in order to avoid disturbing the sonar analysts by the spoken orders given by other operators, headphones must be provided to every analyst and the major part of the human-human communication must be transmitted by headphones. Another possible solution to this problem is to move the consoles away from each other.

5.4 NOISE REDUCTION

Many different types of noises are heard on speech communication channels. Nonspeech-like sounds affect speech intelligibility by masking fragments of speech sounds. To enhance speech obscured by noise, it is necessary to remove the masking sound, or reduce its effect, without distorting the obscured speech sounds. Speech enhancement technology reduces noise and interference in a speech signal regardless of whether the signal is received by telephone, radio, or any other audio source, and independent of the language being spoken. Use of the technology reduces listener fatigue and communication errors.

The Air Force Research Laboratory Information Directorate's Speech Enhancement Unit (SEU) automatically detects and attenuates impulses, tones and wideband random noise using three signal processing algorithms. The impulse algorithm operating in the time domain virtually eliminates all impulse noises, popping and clicking for example. The Digital Spectral Shaping algorithm operating in the frequency domain automatically detects tones and attenuates them to a minimum of 48 dB. The "INTEL" algorithm provides up to 18 dB attenuation of wideband random noise operating in the cepstrum domain. SEU users can select any combination of these noise attenuation processes. Input signals are processed in real time with a maximum system time delay of 300 milliseconds.

Tape recorder noise, receiver noise, wire and radio link noise, automobile ignition noise, powerline hums, and the effects of other interference are reduced by this technology to allow recovery of noise-contaminated conversations. Noise reduction technology reduces communication errors, increases communication range, and improves machine-to-machine communications.

Capabilities

- Real-time reduction of narrowband (tonal) noises
- Real-time reduction of wideband (hissing) noises
- Real-time reduction of impulse (popping, clicking) noises

Benefits

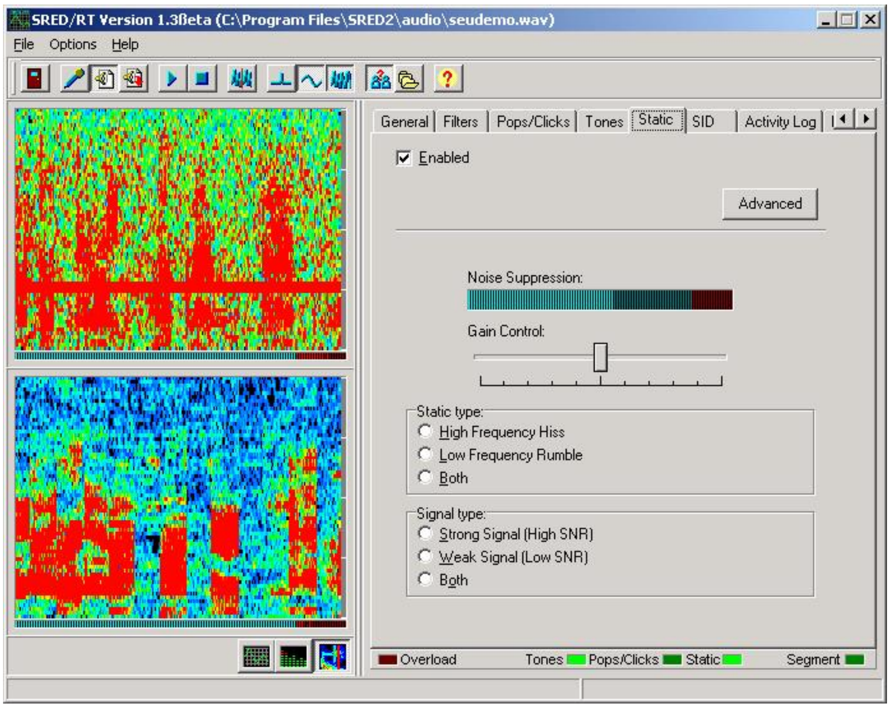
- Reduces communication errors and listener fatigue
- Uncovers signals masked by noise
- Increases communication range
- Improves machine-to-machine communications

Current Applications

- Improved real-time military communications
- Clearer real-time and recorded audio communications for law enforcement activities

Future Applications

- Preprocessor for automatic speech recognition
- Improved hearing aids
- Improved communications range for air traffic control
- Improved radio communications



☑ Enabled

Noise Suppression:

Gain Control:

Static type:
○ High Frequency Hiss
○ Low Frequency Rumble
○ Both

Signal type:
○ Strong Signal (High SNR)
○ Weak Signal (Low SNR)
○ Both

Figure 5.1: The SEU Noise Reduction System.

5.5 TRAINING OF AIR TRAFFIC CONTROLLERS

Air traffic controllers are required to give clear verbal instructions to pilots and are trained to use a very limited English grammar and phraseology. A number of groups have invested significant research effort into finding possible uses for speech recognition technology, both for operational and training use. Many national civil aviation authorities have research programs and several have issued formal specifications for training systems using speech recognition technology. A number of commercial systems are in the advanced stages of development.

At least three potential applications have been identified so far. The simplest application is to monitor the controller's speech for aircraft's callsigns. When identified, flight strip information for the aircraft is activated to allow the controller to inspect or modify it. Since at any time the number of callsigns being used by the controller is quite small, it is possible to build a reasonably reliable system without requiring any special co-operation from the controller. Such systems can provide considerable operational advantages and help reduce the workload on the controller.

A more ambitious use of recognition technology is to help automate the work of a pseudo pilots during training. At present, during training, pseudo pilots listens to the trainee controllers commands and enter the commands into a simulator and give feedback to the controller. During some stages of training more than one pseudo pilot is required per trainee. To provide a fully automatic system requires the recognition of more than a hundred phrases with good recognition accuracy. Systems with smaller phraseology may be used during early stages of training, or for self-study exercises. Since a full training course can require many hundred hours of practice, use of speaker dependent systems is acceptable. In some trials speaker independent systems have provided better results, particularly since the controllers voice can vary significantly during training, or when under stress. While the most obvious advantage of the use of speech recognition technology is to reduce the need for pseudo pilots and so offer more opportunity for practice, incidental advantages include training the controllers to adhere more closely to the textbook phraseology and to adopt a clear and consistent speech style. It is also possible to design systems that automatically log mistakes for later analysis and debriefing.

A third use of recognition technology is when training controllers whose first language is not English. This is a particularly acute problem in Europe. Although such controllers are generally given general purpose English (Weinstein, 1994) language training before their operational training begins, the training often does not give sufficient practice for the highly specific air traffic control vocabulary. A number of multi-media systems incorporating speech recognition technology may be used for phraseology training.

5.6 DARPA SPOKEN LANGUAGE SYSTEMS DEMONSTRATIONS AND APPLICATIONS

Over the past decade, the Defense Advanced Research Projects Agency (DARPA) in the United States has played a crucial role in the research and development of spoken language technology. Tremendous advances have been made in both speech recognition and speech understanding, which have created unprecedented opportunities for major improvements in the effectiveness of human-computer interactions in military, government, and commercial systems. Below are three major programs developing current and future speech technology and applications.

The Translingual Information Detection, Extraction and Summarization (TIDES) program is creating technology to enable English speakers to locate and interpret critical information in multiple languages without requiring knowledge of those languages. The source data could be unformatted raw audio or text, stationary or streaming; critical information could span one or more sources in one or more languages. TIDES technology includes synergistic components for: (i) finding or discovering needed information; (ii) extracting key information about entities, relations, and events; (iii) substantially reducing the amount that a person must read; and (iv) converting foreign language material to English. TIDES has created two text and audio processing systems (known as OnTAP and MiTAP) and is using them in Integrated Feasibility Experiments involving bio-security and terrorism. The experiments, being conducted at contractor facilities with the assistance of military and intelligence personnel, are designed to assess the utility of the evolving technology, to learn where improvements are needed, to develop effective concepts of operation, and to jump-start the transfer of the most effective technology into operational use. Work on Arabic was substantially accelerated in response to the events of September 11. In FY 2003, TIDES will demonstrate initial machine translation capabilities from Chinese and Arabic to English. These demonstrations will be done for Navy and Intelligence Community partners at various U.S. locations. The goal of TIDES is not simply to increase productivity: it provides commanders and other decision-makers with a great deal of timely, vital information that is currently out of reach.

The goal of the DARPA **Communicator** program is to develop and demonstrate “dialogue interaction” technology that enables warriors to talk with computers. Information will be accessible on the battlefield or in command centers without the warfighter ever having to touch a keyboard. The Communicator Platform is wireless and mobile, and will function in a networked environment. Software-enabling dialogue interaction will automatically focus on the context of a dialogue to improve performance. Moreover, the system will adapt to new topics automatically, so that the conversation seems natural and efficient. The technology emphasizes computer-human arbitrated dialogue that uses task knowledge to compensate for natural language effects (e.g., dialects, disfluences, and noisy environments). The majority of the research effort has been on English/computer dialogues in support of command and control operations. Recently, research has begun on foreign language computer interaction in support of coalition operations. Unlike automated translation of news for unlimited vocabulary (speech-to-text, text-to-text) tasks, the effort here is directed toward human-to-machine interactions with task-specific issues that constrain vocabularies. In FY 1999, the program created an open-source architecture for a spoken language dialog system, which is being used by researchers and engineers to experiment with dialogue interaction techniques. In FY 2000, Communicator technology was used for logistic, command and control, and on-the-move information access experiments. DARPA and the sponsoring testers (U.S. Navy and U.S. Marine Corps, through the Small Unit Logistics Advanced Concept Technology Demonstration) evaluated the system and architecture as being highly effective and having potential impact for use in future systems. In FY 2001, hands-on exercises were conducted for small unit logistics operations with the U.S. Marine Corps at Millennium Dragon (using a SINCGARS radio for a field interface) in order to stress-test the technology in extremely noisy and variable environments. In FY 2002, the Communicator system is being stressed in experiments with the Navy on the Sea Shadow and the F/A-18 maintenance mentor at Naval Air Station Patuxent River to support monitoring and alerting of systems, while concurrently improving both information access and distribution. The final Communicator experiment will demonstrate dialogue interaction with a wide array of distributed sensors, heterogeneous databases, and new noisy environments as the U.S. Army evaluates Communicator's ability to automate the combat casualty reporting system. The measure of success will be performance gains for operators using natural dialogue interaction for high-stress and time-critical tasks. Success will validate a new approach for the way 21st century warriors interact with computers, and dialogue interaction will provide for new and effective concepts of operation. A FY 2003 follow-on project focusing Communicator on a command and control problem (e.g., a ship-wide, agent-based dialog network supporting system-wide monitoring and diagnosis aboard the Sea Shadow), as well as a tactical operations task (e.g., fielding the U.S. Army combat casualty system on the Land Warrior platform), may be used to ensure an effective transition mechanism for this revolutionary new interaction technology.

The goal of the **Babylon** program is to develop rapid, two-way, natural language speech translation interfaces and platforms for users in combat and other field environments with constrained military task domains of force protection, refugee processing, and medical triage. The seedling of Babylon, **Rapid Multilingual Support**, was deployed to Afghanistan in the spring of 2002. Also under consideration is the appropriateness of developing a Babylon module for use at Guantanamo Bay, Cuba, to support prisoner interrogation. Babylon will focus on overcoming the many technical and engineering challenges limiting current multilingual translation technology. Babylon will provide an enabling technology to give language support to the warfighter in deciphering possibly critical language communications during operations in foreign territories. The first year (FY 2002) of the Babylon program is to built and rapidly deployed one-way speech translation systems in four target languages – Pashto, Dari, Arabic, and Mandarin – for direct support of overseas field operatives. The systems are delivered in the form of militarized palm-sized PDA devices with 12 hour battery endurance. In FY 2003, each of four Babylon two-way translation teams will develop 10 working-domain-constrained natural language translation prototypes on multiple platforms. Each system will undergo an evaluation process, and the successful teams will advance and continue to refine their systems through technology patches and insertions. In future years, we will expand domains (tasks) supported by our prototypes, and we will improve robustness and enhance the ability of the prototype to meet practical field requirements. This technology is immature and unstable due to the vast complexities of human-to-human communications. Open-domain (multitask), unconstrained dialog translation in multiple environments is still five to 10 years away. DARPA's research is the stimulus to make sure that that capability becomes a reality. Babylon is focusing on low-population, high-terrorist-risk languages that will not be supported by any commercial enterprise.

5.7 BATTLEFIELD BATTLE MANAGEMENT SYSTEM

Between 1997 and 1999, a research programme was carried out at the UK's DERA laboratories to investigate the design of a speech interface for a battlefield battle management system. The system selected for study was the Advanced Land Platform System (ALPS) under development in DERA's Land System group. The work was carried out in conjunction with the Speech Research Unit also within DERA. ALPS provides a prototype system used to investigate the design of the human machine interface (HMI) to and the use of a battlefield battle management system (BBMS). The basic ALPS interface comprises a yoke assembly with central LCD display surrounded by eight soft keys. Also associated with the interface is a numeric keypad, tracker ball, a thumb joystick and various other switches. The system also has three larger LCD multi-function displays which may be used to show a map display and image data from a variety of sensors.

When making a preliminary design of the speech interface, an important decision was the choice between speaker dependent, adaptive or independent recognition technology. Unlike systems for use in air, a system used in a land platform may be used by any of several hundred potential users. Since use of speaker dependent technology requires some mechanism for enrolment and distribution of the templates, it was considered desirable to adopt speaker independent, or speaker adaptive technology if possible.

The staring point for the programme was to determine if speaker independent technology was viable and what size of vocabulary might be deployed. During summer 1998 a series of trials were undertaken to collect a speech database from experienced military commanders in a suitable platform. For reasons of availability a Challenger II tank was selected. Although this would not normally be used as a reconnaissance vehicle, it is tracked and internal noise levels probably equal or exceed any likely to be found in a future platform. Current practice is for the crew to use active noise reduction headsets which incorporate a microphone with voice operated switch. The microphone element is an electret, pressure gradient noise cancelling design. Frequency response is a function of distance of the sound source from the element. For far field sources (such as the engine) the frequency falls at 6dB/octave below 1KHz. For close sound sources, the attenuation is much smaller, but still significant. To optimise the recogniser for the environment a series of recordings were made inside the vehicle and used to provided examples of the types of noise and to evaluate typical signal to noise ratios. A standard speech database was then conditioned to give it the frequency response of the microphone. Models were also trained to match the noise environment of the tank under a variety of conditions. Finally a noise robust algorithm (Parallel Model Combination) was implemented in the recogniser and the models adapted for the expected range of signal to noise ratios.

Trials at Heath Range during 1998 demonstrated the speech interface being used for a simple target selection task. In use an automatic target detection system marked suspect targets on the screen with chevrons. The commander could designate a specific target by speaking a phrase such as ‘select Tango go’. While the word ‘select’ was not essential for the application, it was desirable to include it in the phraseology to ensure that the voice operated switch was fully on and the important word (in this case Tango) was not clipped. The system was demonstrated to a number of military officers and worked satisfactorily both with the tank stationary and in motion.

During 1999 further work was devoted to implementation of a more advanced interface to the ALPS system. The initial design implemented a direct replacement to the tactile interface and used a press to talk switch. For example, to enter a enemy sighting the user might either enter a grid reference from the tracker ball and press buttons to descend through menus, or speak a corresponding phrase such as ‘category tank, type T-80, count 3 confirmed’. To prevent distraction the screen was only updated when the press to talk switch was released. The interface was implemented so that the operator could mix between tactile and speech interface at will. When the press to talk switch was pressed the display showed an indication and showed the available phraseology. The system was trailed on 8 experienced commanders. In general the speech interface was not considered to be particularly useful. Although the speech interface offered a few improvements such as a greater choice of target categories, its design was heavily constrained by the existing menu structure of the interface. A typical comment was: ‘if I have to press a PTT to speak, I might as well press a button to enter the data directly.’ The general conclusion was that when using speech as a direct replacement for a well designed menu based interface offered little improvement.

Before design of the second iteration of the speech interface a series of interviews were held with prospective users. These indicated the following general design criteria:

- Reports are generally of two types, either short formatted messages which contain information for map symbology, or longer reports containing free format speech which is used later during aggregation.
- Speed of entry for a short report is very important; ideally it should be possible to enter a report in 10 seconds.
- The phraseology used in the longer reports is variable and vocabulary is likely to be large.

The final design implemented composed a bimodal interface. The user could toggle between modes through the use of a press switch. In 'assistant mode' (press switch released), the recogniser was placed in a phrase spotting mode and looked for phrases for control of the interface. All such phrases started with the 'callsign' 'Tango Charlie' and could be followed by any of about 30 commands to control map symbology, zoom or to locate a specific callsign. Thus the user cold speak the phrase 'Tango Charlie where is hotel alpha' and the map display would flash the location for unit HA. By pressing the switch the user entered 'report mode' The use then had to start with a phrase for entry of a minimal report which included report type, quantity, type and activity (it was assumed that location information would be entered via tracker ball or laser designator). Thus the user could press and speak 'enemy sighting, three times T-80 moving north'. The speech recognition system extracted the necessary information for map symbology and generated a report which could be manually confirmed and sent with a single key press. As an added refinement the user could speak a short free text message after the formatted information for inclusion as a voice note which might be associated with the report. In addition to use of recognition technology, the system used text to speech to provide users with spoken alerts to new reports from other observers. Through the use of text to speech technology the report could be customised to each observer, for example 'new sighing 5 miles south east of you' rather than 'new sighting at grid 123456'.

In use the system proved highly attractive. The ability to control map symbology allowed user to achieve greater immersion in their observation task than if they had to use a tactile interface. Use of a single press or the press switch allowed a clear demarcation between speech which was 'private' to the observer, and speech which formed part of the report which would be sent externally. It was possible for users to generate minimal reports in less than 5 seconds. Recognition errors were easily corrected either through the tactile interface, or by repeating the phrase. The voice alerts were also found to be useful. An evaluation of use of speech output for confirmation of reports entered was also made, but it was found that unless the confirmation was very succinct, it was highly distracting. The system was generally well liked and operators felt the speech aspects were of value.

In conclusion the work demonstrated that speech interfaces are viable even in the very harsh military environment, but that the interface must be designed with care and not as a direct replacement for a tactile menu based interface.

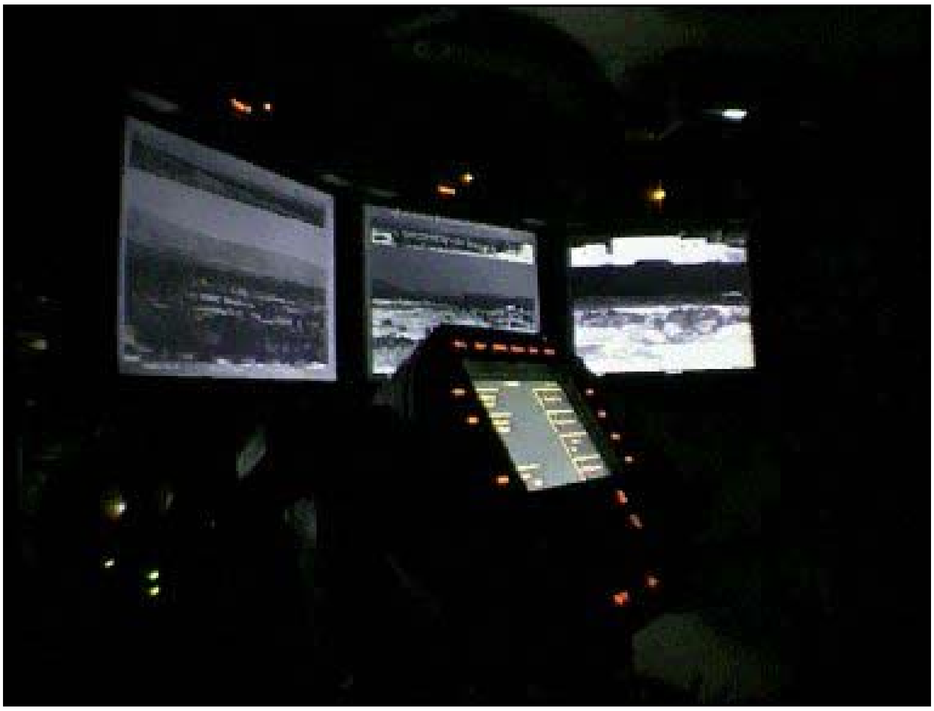


Figure 5.2: The ALPS Display showing a Wide Angle IR Display.

5.8 SPEECH CODERS 600-1200 BPS

The NATO workgroup that is responsible for narrow-band secure-voice coding (AC302/(SG-11)/WG2) has studied the suitability of the new advanced very low bit-rate coders for use in tactical networks. For this purpose, the speech intelligibility of several very low bit-rate speech coding systems was determined (developed in various NATO countries; typical bit rate below 1200 Bps). As a reference, two existing coders were included in the evaluation.

The assessment was performed in four countries: Canada, France, The Netherlands, and the USA. The tests used for this evaluation included Mean Opinion Score tests, CVC-word tests (Consonant-Vowel-Consonant) and Diagnostic Rhyme Tests (DRT). This section presents the results of the experiments performed in The Netherlands and is focused on intelligibility test based on CVC-words. The relation between the CVC-word score, the DRT and the related qualification can be obtained from section 4.2.

The coders, with bit rates ranging from 600-1200 Bps, were evaluated together with two reference coders labeled A and B. The reference coder B is in use in existing 2400 Bps secure voice communication systems.

Coder	Bit rate (Bps)	
A	2400	reference 1
B	2400	reference 2
C	600	
D	800	
E	1200	

As speech-coding systems are normally used in a noisy environment, some of the tests were also performed with additional noise at the input of the coder. In the official assessment, two types of noise at two signal-to-noise ratios were used. This review is limited to one type of noise (speech noise equivalent to voice babble) at a signal-to-noise ratio of 6 dB (about the worst that this type of system can handle). The gender of the speaker was also included as a parameter of the test. In LPC based systems, the intelligibility of female speakers is normally lower than the intelligibility of male speakers. Hence, four test conditions are described here: two signal-to-noise ratios (no noise, and 6 dB) and male and female speech.

The mean CVC-word score (m %) and the standard error (se %) are given in Table 5.1.

Table 5.1: CVC-Word Scores (m %) for Male and Female Speech based upon 16 Speaker-Listener Pairs. The standard errors (se %) are also given.

Coders	Condition							
	Male				Female			
	No noise		SNR 6 dB		No noise		SNR 6 dB	
	M	Se	m	se	m	se	m	se
A	65.1	3.5	33.6	1.3	57.6	2.6	27.6	2.5
B	66.9	2.7	43.0	2.4	65.5	2.4	24.8	1.6
C	48.4	2.3	19.9	1.5	47.9	2.2	17.8	1.3
D	64.0	2.2	28.7	1.5	56.4	1.2	20.7	1.7
E	66.8	2.8	36.0	1.6	54.8	2.5	21.0	1.4

The results indicate that the coders offer a lower intelligibility for female voices. Coders working at a bit rate of 1200 and 800 Bps perform similar to the older 2400 Bps reference systems in the conditions without noise. Finally, noise has a major effect on the performance. Compared with waveform coders and analogue systems, a substantial decrease of the intelligibility is obtained. However, the low bit rate allows a fair transmission under severe jamming conditions. The systems perform at such a level that operational use is foreseen; however improvement of the intelligibility is required.

5.9 CONCLUSION

The primary goal of this report is to describe the military applications of speech and language processing, and the corresponding available technologies.

The military applications are itemized in six categories:

- Command and Control,
- Communications,
- Computers and Information Access,
- Intelligence,
- Training which also includes language training,
- Joint Forces.

For each category a description of the requirements and possible goals are given. The available technologies are subdivided in:

- Speech Processing,
- Language Processing,
- Interaction,
- Assessment and Evaluation.

For these technologies the state-of-the-art with respect to performance and availability is discussed. For speech processing a sub-division for speech coding, speech synthesis and recognition is made.

Also an overview is given of possible assessment procedures and design criteria. Finally some case studies and applications are described.

In brief the reports highlights the need of speech control for operational systems and advanced communications in a changing military environment. Reduction of personnel, increasing complexity of systems, multi-national operations require optimal human performance in which speech can be a natural means of interfacing.

The Research Task Group that performed this study hopes that it will be a useful tool for the Operational staffs, Defense Research Staffs, and potential users within procurement departments of the NATO countries.

Appendix 1 – GLOSSARY OF SPEECH AND LANGUAGE TECHNOLOGY TERMS

Acceptance	Decision outcome, which consists in responding positively to a speaker verification task.
Automatic speech recognition (ASR)	The capability of a machine to convert spoken language to recognized utterances, i.e. the process by which a computer transforms an acoustic speech signal into text.
Automatic translation or machine translation (MT)	Use of a translation system to translate text without human interaction in the actual translation process. The quality of machine-translated text, in terms of terminology, meaning and grammar, varies depending on the nature and complexity of the source text, but is never good enough for publication without extensive editing. – Not to be confused with computer-aided translation!
Computer-aided or computer-assisted translation (CAT)	Translation with the aid of computer programs, such as translation memory tools, designed to reduce the translator's workload and increase consistency of style and terminology. Basically a database in which all previously translated sentences are stored together with the corresponding source text. If, during translation, a sentence appears that is similar to or identical with a previously translated sentence, the program suggests the found target sentence as a possible translation. The translator then decides whether to accept, edit or reject the proposed sentence. – Not to be confused with machine translation!
Continuous speech recognition	A continuous speech system operates on speech in which words are connected together, i.e. not separated by pauses. Continuous speech is more difficult to handle because of a variety of effects. First, it is difficult to find the start and end points of words. Another problem is "coarticulation". The production of each phoneme is affected by the production of surrounding phonemes, and similarly the start and end of words are affected by the preceding and following words. The recognition of continuous speech is also affected by the rate of speech (fast speech tends to be more difficult).
Controlled language	Controlled language (or restricted language) is language, which has been designed to restrict the size of the vocabulary and/or the structure of language used in order to make recognition and processing easier. This is an approach which is particularly valid in certain environments; typical uses of controlled language are in areas where precision of language and speed of response is critical, such as the police and emergency services, aircraft pilots, air traffic control, etc. Controlled language is also used in technical documentation to make the text easier to understand for users or for non-native speakers and to facilitate machine translation.
Dictation system	These systems are usually of the speaker-adaptive type, which means that the user must train the system a certain length of time to adjust its speech recognition functions to the user's speech characteristics. Depending on the quality of the system used, this training will require between thirty minutes and two hours. In the course of training the system software will learn the specific speech of its user and "make a note" regarding his speech profile in a database. Only after this point in time recognition performance will achieve satisfactory results.

Domain	Domain is a term usually applied to the area of application of the language-enabled software, e.g. banking, insurance, travel, medicine, aeronautics, civil engineering, etc. The significance in language engineering is that the vocabulary of an application is often restricted, so the language resource requirements are effectively limited by limiting the domain of application.
False (speaker) acceptance – false (speaker) rejection	Erroneous acceptance of an impostor in speaker detection or in speaker verification. – False rejection is the erroneous rejection of a registered (legitimate) speaker or a genuine speaker in open-set speaker identification or speaker verification.
Hidden Markov Model (HMM)	Hidden Markov Models (HMMs) are commonly used in speech recognition systems to help to determine the words represented by the sound signals captured. An HMM describes the realization of a concatenation of elementary processes which represents the sequence of acoustic parameters extracted from a human utterance.
Impostor	In the context of speaker identification, an impostor is an applicant speaker who does not belong to the set of registered speakers. In the context of speaker verification, an impostor is a speaker whose real identity is different from his claimed identity. Alternative terms: impersonator, usurpator (both are very rarely used).
Intelligibility	Measure of the proportion of speech that is understood. It is usually quantified as the percentage of a message understood correctly.
Isolated speech (or isolated word) recognition	An isolated-word recognition system operates on single words at a time, requiring a distinct pause between saying each word. This is the simplest form of recognition to perform because the end points are easier to find and the pronunciation of a word tends not to affect others. Thus, because the occurrences of words are more consistent, they are easier to recognize. – The technique of this system is to compare the incoming speech signal with an internal representation of the acoustic pattern of each word in a relatively small vocabulary and to select the best match, using some distance metric.
Language engineering	Language engineering is the application of knowledge of language to the development of computer systems, which can recognize, understand, interpret and generate human language in all its forms.
Language identification	Language identification aims at predicting the likely language from an unknown speech signal.
Large (very-large, medium or small) vocabulary speech recognition	The size of the vocabulary of a speech recognition system affects the complexity, processing requirements and the accuracy of the system. Some applications only require a few words (e.g. numbers only), others require very large dictionaries (e.g. dictation systems). There are no fixed definitions with respect to the various size categories, however, the following grouping may be acceptable for practical purposes: • very-large vocabulary – tens of thousands of words; • large vocabulary – thousands of words; • medium vocabulary – hundreds to thousand of words; • small vocabulary – tens of words.
Lombard effect	The effect that occurs when humans speak at a higher level (use more vocal effort) in conditions of higher environmental noise.
Multi-lingual	The adjective multi-lingual is properly used to mean that something exists in a form that can handle several languages but is, in practice, often used to describe the characteristic that versions exist for several languages.

Multi-speaker detection (or open-set speaker identification)	Multi-speaker detection is used in applications where one first finds messages spoken by any of a number of speakers of interest, and then further identifies which of the speakers of interest is speaking.
Natural language	A human language, as distinct from computer languages, which are used to write the programming codes that make software activate hardware to perform a variety of tasks. Any non-invented language is a natural language.
Natural language processing (NLP)	Natural language processing is a term in use since the 1980s to define a class of software which handles texts, written in natural language, intelligently. NLP software includes word processors, dictionaries, grammar and spell checkers, and translation programs.
Non-native language; non-native speaker	Any language that is not one's mother tongue, i.e. a language not learnt during early childhood. – A speaker making utterances in a language which is not his mother tongue.
Open-set speaker identification	This concept is used in scenarios where the speaker identification system has the ability to recognize when an unknown speech sample has come from a new speaker. If the unknown sample has come from a known speaker the system reports this recognized speaker.
Phoneme	The smallest meaningful sound unit of speech which can be identified from an acoustic flow of speech. In the English language, for instance, there are about 40 different basic phonemes.
Prosody	Those properties of speech utterances that cannot be derived in a straightforward fashion from the identity of the vowel and consonant phonemes that are strung together in the linguistic representation underlying the speech utterance, e.g. intonation (i.e. speech melody), word and phrase boundaries, (word) stress, (sentence) accent, tempo, and changes in speaking rate.
PSOLA	Pitch synchronous overlap-add technique used in concatenative speech synthesis.
Signal-to-noise ratio (SNR)	In general signal description terms, the ratio of the amplitude of the desired signal to the amplitude of noise signals at a given point in time. In automatic speech recognition, the ratio of information-carrying signals (speech signals) to background noise. Usually expressed in decibels (dB).
Source language	The language of a text to be translated. Accordingly, the language into which the text is to be translated is called the target language.
Speaker detection	Speaker detection is conceptually similar to the speaker verification problem. The task is to sort a set of speech utterances by the likelihood that a particular speaker of interest speaks each.
Speaker identification	A decision-making procedure by which the speech sample of a speaker is compared with a group of speech samples. It is assumed that the test person's speech sample is contained in the group. The task of speaker identification consists in identifying that speech sample which shows the greatest similarity with the speaker's sample; in other words, speaker identification is the capability to identify a speaker from a group of speakers. The error rate rises with the increasing number of speech samples within the group.
Speaker recognition	Speaker recognition is the process of automatically recognizing who is speaking, on the basis of individual information included in speech signals. It can be divided into speaker identification and speaker verification.

Speaker verification	A decision-making procedure by which the utterance of a person is compared with a speech sample of the same person which has been deposited before. If the degree of correspondence between the utterance and the sample exceeds a certain threshold, the person is accepted. Hence, speaker verification is a method of confirming that a speaker is the same person he or she claims to be.
Speaker-dependent, speaker-adaptive and speaker-independent recognition systems	A speaker-dependent system is developed to operate for a single speaker. These systems are usually easier to develop, cheaper to buy and more accurate, but not as flexible as speaker-adaptive or speaker-independent systems. A speaker-adaptive system is developed to adapt its operation to the characteristics of new (additional) speakers. Its difficulty lies somewhere between speaker-independent and speaker-dependent systems. A speaker-independent system is developed to recognize speech regardless of the speaker, i.e. it does not need to be trained to recognize individual speakers. These systems are the most difficult to develop, most expensive, and accuracy is lower than with speaker-dependent systems. However, they are more flexible.
Speech coding	Speech coding aims at encoding speech in a digital format that can be transmitted over digital links and then decoded at the receiving end. When in digital form, it is naturally very easy to encrypt. Additionally, speech coding usually aims at producing a compact representation of speech sounds such that when reconstructed it is perceived to be close to the original. The two main measures of closeness are intelligibility and naturalness.
Speech communicability	Measure of the ease with which speech communication is performed. It includes speech intelligibility, speech quality, vocal effort, and delays.
Speech communication	Conveying or exchanging information using speech, speaking, and hearing modalities. Speech communication may involve brief texts, sentences, groups of words and isolated words.
Speech corpus (or spoken language corpus)	Any collection of speech recordings which is accessible in computer-readable form and which comes with annotation and documentation sufficient to allow re-use. For example, a corpus could comprise recordings of car drivers speaking to a simulation of a voice-operated control system which recognizes spoken commands. Such a corpus is then used to help establish the user requirements for a voice-operated control system for the market. Speech corpora may be monolingual or multi-lingual.
Speech output system	Some artifact, either a dedicated machine or a computer program that produces signals that are intended to be functionally equivalent to speech produced by humans.
Speech synthesis	Speech synthesis is the name given to the production of speech sounds by a machine. Computer programs convert written input to spoken output. This process is often referred to as text-to-speech conversion.
Speech understanding	The use of artificial intelligence techniques to process and interpret audio signals representing human speech. This topic is still very much a research issue, but some advances have been made in the field.
Speech-to-Text system	A system that converts the acoustic signals of spoken language into written text. A typical example of speech-to-text application is in dictation systems. Speech-to-text systems may also be used as an intermediate step for making spoken language accessible to machine translation.

Spontaneous speech	Spontaneous speech is often used synonymously with continuous speech but more explicitly recognizing that there are other characteristics of speech which make it difficult to understand, such as the tendency for people not to speak grammatically correctly or to speak in ways which make it difficult to maintain consistent context.
Summarization	The process and result of producing a concise description of a document, which covers the full scope of its contents.
Target language	The language into which a text is to be translated. Accordingly, the text from which the text is translated is called the source language.
Text	In speech and language technology, the term text is used frequently to distinguish written, printed, or symbolically recorded (using character encoding) language from speech.
Text-to-Speech system	A speech output system that converts written text (generally stored in a computer memory in ASCII code) into speech.
Topic spotting	Topic spotting is a term used to describe a number of techniques used to monitor a stream of messages (text or speech) for those of particular interest. Typical applications include automatic surveillance and screening of messages before referring to human operators.
Translation	The act and result of rendering written text from one language into another.
Translation Memory (TM)	Computer-aided translation program. In essence a database that stores translated sentences (translation units or segments) with their respective source segments. For each new segment to be translated, the program scans the database for a previous source segment that matches the new segment exactly or approximately (fuzzy match) and, if found, suggests the corresponding target segment as a possible translation. The translator can then accept, modify or reject the suggested translation.
Utterance	A spoken phrase or passage; in a speech recognition context, the string of sounds produced by a speaker between two pauses.
Voice characteristics	The aspects of speech brought about by individual features of a human voice. These features are unique to a person and can help speech-recognizing systems in identifying a speaker.
Wizard-of-Oz testing	Testing in which the behavior of an interactive automation is simulated by a human being, but in such a way that the person participating in the test or experiment is unaware of the substitution.
Word error rate	The fraction of errors made by a recognition system, i.e. the number of errors divided by the number of words to be recognized. Often expressed as a percentage.
Word spotting	Word spotting is a term used to describe a number of techniques used to monitor a stream of spoken messages for specific individual words of particular interest. This is an easier task than topic spotting, which also involves associating words with certain topics of interest.

Appendix 2 – LIST OF AUTHORS ON ORIGINAL REPORT

P. Alinàt	Thomson Sintra ASM, France
M. Bates	BBN Systems and Technologies, USA
S. Bodenkamp	Amt für Auslandsfragen, Germany
A.W. Bronkhorst	TNO Human Factors Research Institute, The Netherlands
E.J. Cupples	Rome Laboratory/IRAA, USA
L. Gagnon	Communications Security Establishment, Dept. of Nat. Defence, Canada
F.F. Leyendecker	Amt fur Nachrichtenwesen der Bunderwehr Abt. I, Germany
D.A. van Leeuwen	TNO Human Factors Research Institute, The Netherlands
R. Martinez	Ministerio de Defensa, Spain
R.K. Moore	Speech Research Unit DRA, UK
G. O'Leary	MIT Lincoln Laboratory, USA
J.M. Pardo	Universidad Politécnica de Madrid, Department of Electronic Engineering, Spain
J. Payette	Canadian Space Agency, Canada
E.W. Pijpers	National Aerospace Laboratory, The Netherlands
Chr. Rouchouze	Delegation Générale pour l'Armement/ Direction des Constructions Navales, France
A.M. Schaafstal	TNO Human Factors Research Institute, The Netherlands
R.W. Series	Speech Research Unit DRA, UK
H.J.M. Steeneken	TNO Human Factors Research Institute, The Netherlands
A.J. South	DRA Air Systems, UK
C. Swail	Institute for Aerospace Research, National Research Council, Canada
M.M. Taylor	DCIEM, Canada
I.M. Trancoso	Institutio de Engenharia de Sistemas e Computadores, Instutio Superior Técnico, Portugal
Ph. Valéry	Sextant Avionique, France
C.R.A. Vloeberghs	Royal Military Academy Brussels, Belgium
C.J. Weinstein	MIT Lincoln Laboratory, USA
D. Windheiser	Delegation Générale pour l'Armement/Direction de la Recherche et de la Technologie, France

Appendix 3 – LIST OF AUTHORS FOR UPDATED REPORT

T. Anderson	Air Force Research Laboratory, USA
C. Bruckner	Bundessprachenmant, Germany
P. Collins	DSTL, UK
E. Geoffrois	DGA-CTA/GIP, France
J. Grieco	Air Forces Research Laboratory, USA
D.A. van Leeuwen	TNO Human Factors Research Institute, The Netherlands
O.D. Orman	TUBITAK-UEKAE, Turkey
H. Palaz	TUBITAK-UEKAE, Turkey
S. Pigeon	Royal Military Academy Brussels, Belgium
R.W. Series	20/20 Speech LTD, UK
H.J.M. Steeneken	TNO Human Factors Research Institute, The Netherlands
A.J. South	20/20 Speech LTD, UK
H. Stumpf	Bundessprachenmant, Germany
C. Swail	Institute for Aerospace Research, National Research Council, Canada
A. Teixeira	IEETA, Portugal
C. Teixeira	INESC ID, Portugal
S.J. van Wijngaarden	TNO Human Factors Research Institute, The Netherlands
M. Zissman	MIT Lincoln Laboratory, USA

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
	RTO-TR-IST-037 AC/323(IST-037)TP/22	ISBN 92-837-1144-0	UNCLASSIFIED/ UNLIMITED

5. Originator

Research and Technology Organisation
North Atlantic Treaty Organisation
BP 25, F-92201 Neuilly-sur-Seine Cedex, France

6. Title

Use of Speech and Language Technology in Military Environments

7. Presented at/Sponsored by

The Information Systems Technology (IST) to support a Lecture Series presented on 20-21 November 2003 in Montreal, Canada; 1-2 December 2003 in Arcueil, France; and 4-5 December 2003 in Istanbul, Turkey.

8. Author(s)/Editor(s)	9. Date
Multiple	December 2005
10. Author's/Editor's Address	**11. Pages**
Multiple	106

12. Distribution Statement

There are no restrictions on the distribution of this document. Information about the availability of this and other RTO unclassified publications is given on the back cover.

13. Keywords/Descriptors

Automatic speech recognition
Control equipment
Human factors engineering
Identification systems
Intelligibility
Language identification
Language recognition
Languages
Linguistics
Man computer interface
Man machine systems
Methodologies
Multilingualism
Performance tests
Reviews
Speaker recognition
Speech recognition
Speech technology
Standards
Voice communication
Voice control
Voice recognition
Word association

14. Abstract

Communications, command and control, intelligence and training systems are making more and more use of speech and language technology components: i.e. speech coders, voice controlled C2 systems, speaker and language recognition, translation systems and automated training suites. Implementation of these technologies requires an understanding of what performance is possible with the products that are available today and those that will likely to be available in the next few years.

As speech and language technology become more available for integration into military systems, it is important that those involved in system design and program management be aware of the capabilities and the limitations of present speech systems. They must also be aware of the current state of research in order to be able to consider what will be possible in the future. This will be very important when considering future military systems upgrades.

This lecture series includes presentations of the current state of the art and the current research topics in selected speech and language technology areas: assessment techniques and standards, speech recognition, speaker and language identification, and translation.

NORTH ATLANTIC TREATY ORGANISATION

NATO
OTAN

BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

RESEARCH AND TECHNOLOGY ORGANISATION

DIFFUSION DES PUBLICATIONS
RTO NON CLASSIFIEES

Les publications de l'AGARD et de la RTO peuvent parfois être obtenues auprès des centres nationaux de distribution indiqués ci-dessous. Si vous souhaitez recevoir toutes les publications de la RTO, ou simplement celles qui concernent certains Panels, vous pouvez demander d'être inclus soit à titre personnel, soit au nom de votre organisation, sur la liste d'envoi.

Les publications de la RTO et de l'AGARD sont également en vente auprès des agences de vente indiquées ci-dessous.

Les demandes de documents RTO ou AGARD doivent comporter la dénomination « RTO » ou « AGARD » selon le cas, suivi du numéro de série. Des informations analogues, telles que le titre est la date de publication sont souhaitables.

Si vous souhaitez recevoir une notification électronique de la disponibilité des rapports de la RTO au fur et à mesure de leur publication, vous pouvez consulter notre site Web (www.rta.nato.int) et vous abonner à ce service.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Streitkräfteamt / Abteilung III
Fachinformationszentrum der
Bundeswehr (FIZBw)
Friedrich-Ebert-Allee 34, D-53113 Bonn

BELGIQUE
Etat-Major de la Défense
Département d'Etat-Major Stratégie
ACOS-STRAT – Coord. RTO
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
DSIGRD2
Bibliothécaire des ressources du savoir
R et D pour la défense Canada
Ministère de la Défense nationale
305, rue Rideau, 9e étage
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Defence Research Establishment
Ryvangs Allé 1, P.O. Box 2715
DK-2100 Copenhagen Ø

ESPAGNE
SDG TECEN / DGAM
C/ Arturo Soria 289
Madrid 28033

ETATS-UNIS
NASA Center for AeroSpace
Information (CASI)
Parkway Center, 7121 Standard Drive
Hanover, MD 21076-1320

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72, 92322 Châtillon Cedex

GRECE (Correspondant)
Defence Industry & Research
General Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HONGRIE
Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ISLANDE
Director of Aviation
c/o Flugrad
Reykjavik

ITALIE
Centro di Documentazione
Tecnico-Scientifica della Difesa
Via XX Settembre 123
00187 Roma

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

PAYS-BAS
Royal Netherlands Military
Academy Library
P.O. Box 90.002
4800 PA Breda

POLOGNE
Armament Policy Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide
P-2720 Amadora

REPUBLIQUE TCHEQUE
LOM PRAHA s. p.
o. z. VTÚLaPVO
Mladoboleslavská 944
PO Box 18
197 21 Praha 9

ROYAUME-UNI
Dstl Knowledge Services
Information Centre, Building 247
Dstl Porton Down
Salisbury
Wiltshire SP4 0JQ

TURQUIE
Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi Başkanlığı
06650 Bakanliklar – Ankara

AGENCES DE VENTE

NASA Center for AeroSpace
Information (CASI)
Parkway Center, 7121 Standard Drive
Hanover, MD 21076-1320
ETATS-UNIS

The British Library Document
Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
ROYAUME-UNI

Canada Institute for Scientific and
Technical Information (CISTI)
National Research Council
Acquisitions, Montreal Road, Building M-55
Ottawa K1A 0S2, CANADA

Les demandes de documents RTO ou AGARD doivent comporter la dénomination « RTO » ou « AGARD » selon le cas, suivie du numéro de série (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Des références bibliographiques complètes ainsi que des résumés des publications RTO et AGARD figurent dans les journaux suivants :

Scientific and Technical Aerospace Reports (STAR)
STAR peut être consulté en ligne au localisateur de
ressources uniformes (URL) suivant:
http://www.sti.nasa.gov/Pubs/star/Star.html
STAR est édité par CASI dans le cadre du programme
NASA d'information scientifique et technique (STI)
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
ETATS-UNIS

Government Reports Announcements & Index (GRA&I)
publié par le National Technical Information Service
Springfield
Virginia 2216
ETATS-UNIS
(accessible également en mode interactif dans la base de
données bibliographiques en ligne du NTIS, et sur CD-ROM)

NORTH ATLANTIC TREATY ORGANISATION

RESEARCH AND TECHNOLOGY ORGANISATION

BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DISTRIBUTION OF UNCLASSIFIED RTO PUBLICATIONS

AGARD & RTO publications are sometimes available from the National Distribution Centres listed below. If you wish to receive all RTO reports, or just those relating to one or more specific RTO Panels, they may be willing to include you (or your Organisation) in their distribution.

RTO and AGARD reports may also be purchased from the Sales Agencies listed below.

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number. Collateral information such as title and publication date is desirable.

If you wish to receive electronic notification of RTO reports as they are published, please visit our website (www.rta.nato.int) from where you can register for this service.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Etat-Major de la Défense
Département d'Etat-Major Stratégie
ACOS-STRAT – Coord. RTO
Quartier Reine Elisabeth
Rue d'Evère
B-1140 Bruxelles

CANADA
DRDKIM2
Knowledge Resources Librarian
Defence R&D Canada
Department of National Defence
305 Rideau Street
9th Floor
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC
LOM PRAHA s. p.
o. z. VTÚLaPVO
Mladoboleslavská 944
PO Box 18
197 21 Praha 9

DENMARK
Danish Defence Research
Establishment
Ryvangs Allé 1
P.O. Box 2715
DK-2100 Copenhagen Ø

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72
92322 Châtillon Cedex

GERMANY
Streitkräfteamt / Abteilung III
Fachinformationszentrum der
Bundeswehr (FIZBw)
Friedrich-Ebert-Allee 34
D-53113 Bonn

GREECE (Point of Contact)
Defence Industry & Research
General Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HUNGARY
Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ICELAND
Director of Aviation
c/o Flugrad, Reykjavik

ITALY
Centro di Documentazione
Tecnico-Scientifica della Difesa
Via XX Settembre 123
00187 Roma

LUXEMBOURG
See Belgium

NETHERLANDS
Royal Netherlands Military
Academy Library
P.O. Box 90.002
4800 PA Breda

NORWAY
Norwegian Defence Research
Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

POLAND
Armament Policy Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide, P-2720 Amadora

SPAIN
SDG TECEN / DGAM
C/ Arturo Soria 289
Madrid 28033

TURKEY
Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi Başkanlığı
06650 Bakanliklar – Ankara

UNITED KINGDOM
Dstl Knowledge Services
Information Centre, Building 247
Dstl Porton Down
Salisbury, Wiltshire SP4 0JQ

UNITED STATES
NASA Center for AeroSpace
Information (CASI)
Parkway Center, 7121 Standard Drive
Hanover, MD 21076-1320

SALES AGENCIES

NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320
UNITED STATES

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
UNITED KINGDOM

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council
Acquisitions
Montreal Road, Building M-55
Ottawa K1A 0S2, CANADA

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Full bibliographical references and abstracts of RTO and AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
STAR is available on-line at the following uniform
resource locator:
http://www.sti.nasa.gov/Pubs/star/Star.html
STAR is published by CASI for the NASA Scientific
and Technical Information (STI) Program
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
UNITED STATES

Government Reports Announcements & Index (GRA&I)
published by the National Technical Information Service
Springfield
Virginia 2216
UNITED STATES
(also available online in the NTIS Bibliographic
Database or on CD-ROM)

ISBN 92-837-1144-0