RTO-TR-10
AC/323(IST)TP/5

RTO-TR-10

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION

BP 25, 7 RUE ANCELLE, F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

RTO TECHNICAL REPORT 10

The Impact of Speech Under "Stress" on Military Speech Technology

(l'Impact de la parole en condition de "stress" sur les technologies vocales militaires)

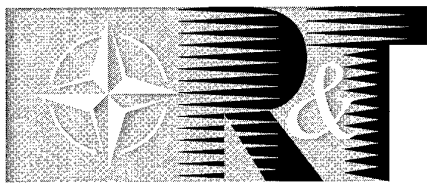
This Technical Report has been prepared as a result of a project on "Speech under Stress Conditions" for the RTO Information Systems Technology Panel (IST).

Published March 2000

Distribution and Availability on Back Cover

RTO-TR-10
AC/323(IST)TP/5

NORTH ATLANTIC TREATY ORGANIZATION



RESEARCH AND TECHNOLOGY ORGANIZATION

BP 25, 7 RUE ANCELLE, F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

RTO TECHNICAL REPORT 10

The Impact of Speech Under "Stress" on Military Speech Technology

(l'Impact de la parole en condition de "stress" sur les technologies vocales militaires)

by

Prof. Claude VLOEBERGHS, Mr. Patrick VERLINDE, Belgium
Mr. Carl SWAIL, Canada
Dr Herman STEENEKEN (Chairman), Dr. David van LEEUWEN, The Netherlands
Prof. Isabel TRANCOSO (Secretary), Portugal
Mr. Allan SOUTH, Prof. Roger MOORE, United Kingdom
Mr. E. James CUPPLES, Dr. Timothy ANDERSON, Prof. John HANSEN, USA

This Technical Report has been prepared as a result of a project on "Speech under Stress Conditions" for the RTO Information Systems Technology Panel (IST).

The Research and Technology Organization (RTO) of NATO

RTO is the single focus in NATO for Defence Research and Technology activities. Its mission is to conduct and promote cooperative 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 coordination 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 coordinates RTO's cooperation 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 initial cooperation.

The total spectrum of R&T activities is covered by 7 Panels, dealing with:

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

These Panels are made up of national representatives as well as generally recognised 'world class' scientists. The Panels 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 cooperation 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.

Printed on recycled paper

Published March 2000

ISBN 92-837-1027-4

Printed by Canada Communication Group Inc.
(A St. Joseph Corporation Company)
45 Sacré-Cœur Blvd., Hull (Québec), Canada K1A 0S7

The Impact of Speech Under "Stress" on Military Speech Technology

(RTO TR-10)

Executive Summary

The field of military speech technology requires the integrated use of speech systems for communications, command, and control and intelligence. Speech technology in military environments offers the promise of more direct and effective communication, speaker verification of personnel, and allowing operators to have access to better information. The problems of battlefield stress conditions, however, raise a serious obstacle for the transition of commercial off-the-shelf speech technology for speech recognition, speaker verification, synthesis and coding. Studies conducted by participating NATO laboratories and discussed here suggest that many COTS speech systems which were designed for quiet or low-noise office environments, cannot be effectively used in real-world, high task stress, emotional induced, high background noise, and operator fatigued situations. The main findings and recommendations are:

- Military operations are often conducted under conditions of stress induced by high workload, sleep deprivation, fear and emotion, confusion due to conflicting information, psychological tension, pain, and other typical conditions encountered in the modern battlefield context. These conditions are known to affect the physical and cognitive abilities of human speech characteristics.
- It is suggested that the effect of operator based stress factors on speech production quality is likely to be detrimental to the effectiveness of communication in general, in particular to the performance of communication equipment and weapon systems equipped with vocal interfaces (e.g., advanced cockpits, command, control, and communication systems, information warfare).
- Commercial off-the-shelf speech recognition systems are not yet able to address the wide speaker variability associated with speech produced under stress.
- Databases obtained or collected during this study have been distributed to all participating NATO countries, and most are available in CD-ROM format.
- Progress in the field of military based speech technology, including advances in speech based system design, has been restricted due to the lack of available databases of speech under stress. In particular, the type of stress that an operator may experience in the modern battlefield context is not easily simulated, and therefore it is difficult to systematically collect speech data for use in research and speech system training.
- In the future it will be more necessary to improve the coordination of multi-national military forces. The need therefore exists for battlefield simulations with multi-national military personnel using a wide range of speech technology. Such battlefield simulations will have to address the impact of factors such as high workload, sleep deprivation, fear and emotion, confusion, psychological tension, pain, etc. on speech technology.

L'impact de la parole en condition de "stress" sur les technologies vocales militaires

(RTO TR-10)

Synthèse

Le domaine des technologies vocales militaires concerne l'intégration de systèmes de parole pour les communications, le commandement et contrôle et le renseignement. La mise en œuvre des technologies vocales dans des environnements militaires ouvre la perspective de communications plus directes et plus efficaces, avec vérification du locuteur, permettant aux opérateurs d'accéder à des informations plus fiables. Cependant, les problèmes occasionnés par le stress du champ de bataille représentent un obstacle sérieux à la transition des technologies vocales disponibles sur étagère (COTS) vers des applications militaires de reconnaissance de la parole, de vérification du locuteur, de synthèse et de codage. Les études réalisées en coopération par des laboratoires de l'OTAN indiquent que bon nombre de systèmes de parole, conçues pour des environnements de bureau peu bruyants, sont inadaptés à des situations réelles, à haut niveau de stress opérationnel et d'émotion, avec des niveaux de bruit de fond élevés, impliquant des opérateurs fatigués. Les principales conclusions et recommandations sont les suivantes:

- Les opérations militaires sont souvent conduites dans des conditions de stress induites par des charges de travail élevées, le manque de sommeil, la peur et l'émotion, la confusion due à des informations contradictoires, la tension psychologique, la douleur, et par d'autres conditions typiques du champ de bataille moderne. Il a été démontré que ces conditions affectent les capacités physiques et cognitives des caractéristiques de la parole humaine.
- Il est soutenu que l'effet des éléments stressants sur la qualité de la parole risque de nuire à l'efficacité de la communication en général, en particulier en ce qui concerne les performances du matériel de communication et des systèmes d'armes équipés d'interfaces vocales (par exemple les postes de pilotage avancés, les systèmes de commandement, contrôle et communications et de guerre de l'information).
- Les systèmes de reconnaissance de la parole disponibles sur étagère ne sont pas encore en mesure de gérer les grandes variations entre locuteurs associées à la production de la parole sous le stress.
- Des bases de données obtenues ou recueillies au cours de cette étude ont été diffusées à l'ensemble des pays de l'OTAN participants, et la plupart d'entre elles sont disponibles sous forme de CD-ROM.
- Les avancées dans le domaine des technologies vocales militaires, y compris les avancées dans la conception des systèmes de parole, ont été freinées par la non-disponibilité de bases de données contenant des exemples de paroles produites sous le stress. En particulier, le type de stress éprouvé par un opérateur dans le contexte du champ de bataille moderne n'est pas facile à simuler. Par conséquent, il est difficile de collecter de façon systématique des données vocales pour incorporation aux programmes de formation aux systèmes de parole et pour la recherche.
- A l'avenir, il deviendra de plus en plus nécessaire d'améliorer la coordination des forces militaires internationales. Le besoin existe donc, de simulations du champ de bataille intégrant des personnels militaires internationaux et faisant appel à un éventail de technologies vocales. De telles simulations du champ de bataille devront tenir compte de l'impact de facteurs tels que la charge de travail élevée, le manque de sommeil, la peur et l'émotion, la confusion, la tension psychologique, la douleur etc. sur les technologies vocales.

Contents

Preface

Military operations are often conducted under conditions of stress, induced by high workload, sleep deprivation and, battle stress. These stresses are believed to affect voice quality, and are likely to be detrimental to the performance of communication equipment (e.g. low-bitrate secure voice systems) and weaponry with vocal interfaces (e.g., advanced cockpits, command, and control systems). The actual effects of stress on voice are not well understood. IST Task Group 001 (former RSG.10) has conducted a study on stress effects of the kind to which military operations are subject. The work was separated into five tasks:

1. Collect speech data for various types of stress, such as for workload. In parallel stress related physiological measures and objective measures will be collected,
2. Produce an annotated database that might be used beyond the confines of the Task Group (continuous data base collection through life time of the project),
3. Characterise speech parameters related to stress,
4. Assess effects on performance of recognisers and communication equipment,
5. Relate derived results to military applications.

In this report the results of the study are presented. These results were also presented and discussed at a special session of the International Conference on Acoustics, Speech and Signal Processing held in 1999 at the Phoenix USA Conference Center under responsibility of the IEEE and the IST-011/TG-001.

Préface

Les opérations militaires sont souvent conduites en conditions de stress, du fait de la charge de travail élevée, du manque de sommeil ou du stress au combat. Ces stress affectent la qualité de la voix et peuvent diminuer la performance des équipements de communication (par exemple les systèmes de communication sécurisés à bas débit) et les armements à interface vocale (par exemple les systèmes avancés de "copilote électronique", de commande et de contrôle). Les effets réels du stress sur la voix ne sont pas connus précisément. Le groupe IST-011/TG-001 (ex-RSG.10) a conduit une étude sur les effets de stress du type de ceux que l'on rencontre en opérations militaires. Le travail a été réparti en cinq tâches :

1. Collecter des données de parole sous différents types de stress, comme la charge de travail, ainsi que des mesures physiologiques corrélées à d'autres mesures objectives,
2. Produire une base de donnée annotée qui pourra être utilisée au-delà du seul groupe OTAN (perrenité du projet),
3. Caractériser les paramètres de la voix liés au stress,
4. Evaluer l'impact sur la performance des systèmes de reconnaissance de la parole et de communication,
5. Applications militaires des résultats.

Le rapport présente les résultats de l'étude. Ces résultats ont aussi été présentés dans une session spéciale du congrès international ICASSP (International Conference on Acoustics, Speech and Signal Processing) qui s'est tenue en 1999 à Phoenix sous la responsabilité conjointe de l'IEEE et du groupe IST-011/TG-001.

Foreword

Efficient speech communication is recognized as a critical and instrumental capability in many military applications such as command and control, aircraft and vehicle operations, military communication, translation, intelligence, and training. The former NATO research study group on speech processing (AC243 (Panel 3)RSG.10) conducts since its establishment in 1978 experiments and surveys focused on military applications of language processing. Guided by its mandate, the former RSG.10 initiated in the past the publication of overviews on potential applications of speech technology for military use and also organized several workshops and lecture series on military-relevant speech technology topics. Recently the group continued under the IST panel as AC232/IST/TG001.

In recent years, the speech R&D community has developed or enhanced many 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 control and command environments to substantially reduce the workload of operators. In many situations operators have busy eyes and hands, and must use other media such as speech to control functions and receive feedback messages;
- Large vocabulary speech recognition and speech understanding systems are useful as training aid and to prepare for missions;
- Speech processing techniques are available to identify talkers, languages, and keywords and can be integrated into military intelligence systems;
- Automatic training systems combining automatic speech recognition and synthesis technologies can be utilized to train personnel with minimum or no instructor participation (e.g. Air traffic controllers).

This report is the result of a project on "Speech under Stress Conditions" with contributions of all Task Group members which represent nine NATO countries (Belgium, Canada, France, Germany, the Netherlands, Portugal, Spain, United Kingdom, and the United States; in 1999 Turkey joined this group).

Because speech technologies are constantly improving and adapting to new requirements, it is the intention of the Task Group to initiate projects on military applications of speech technology. Therefore the group appreciates any comment and feedback on this report.

Membership of Information System Technology Task Group 001 "Speech and Language Technology"

Chairman
Dr. Herman J.M. Steeneken
TNO Human Factors Research Institute
P.O. Box 23
3769 ZG Soesterberg
The Netherlands

Secretary
Prof. Isabel Trancoso
INESC, Speech Processing Group
R. Alves Redol, 9
1000 Lisbon
Portugal

Members

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

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

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

Prof. John H.L. Hansen
Center for Spoken Language Understanding
Box 258, Univ. of Colorado at Boulder
Boulder, Colorado 80309-0258
USA

Prof. Jean Paul Haton
Universite Henri Poincaré
LORIA B.P. 239
54506 Vandoevre-les-Nancy
France

Mr. Rafael Martinez
Ministerio de Defensa
Av. Padre Huidobro, km 8,500
28023 Madrid
Spain

Prof. Roger K. Moore
Speech Research Unit
DERA Malvern, St. Andrews Road
Great Malvern, Worcs WR14 3PS
United Kingdom

Mr. Hasan Palaz
TUBITAK-AEKAE, National Research Institute
of Electronics & Cryptology
P.K. 21, 41470 Gebze. Kocaeli
Turkey

Prof. José M. Pardo
ETSI de Telecomunicacion - UPM
Ciudad Universitaria
28040 Madrid
Spain

Dr. Dough Reynolds
Information Systems Technology Group
MIT Lincoln Laboratory, 244 Wood Street
Lexington, MA 02420-9108
USA

Mr. Alan J. South
DERA Farnborough
System Integration Dept.
Room 2067, Probert (A5) Building
Farnborough, Hants GU14 0LX
United Kingdom

Mr. H. Stumpf
Bundessprachenamt
Horbeller Strasse 52
50354 Huerth
Germany

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

Maj. Patrick Verlinde
Royal Military Academy
Renaissancelaan 30
B-1000 Brussels
Belgium

Panel Executive
Lt. Col. Alain Gouay
RTA/IST
BP 25
7, rue Ancelle
F-92201 Neuilly-sur-Seine Cedex
France

Chapter 1

Introduction

Military operations are often conducted under conditions of physical and mental stress which are detrimental to the effectiveness of speech communications. This stress is induced by high workload, sleep deprivation, frustration over contradictory information, emotions such as fear, pain, psychological tension, and other modern battlefield conditions. These conditions are known to affect the physical and cognitive production of human speech. The change in speech production is likely to be disastrous not only to human understanding of coded speech, but also to the performance of communication equipment, and weapon systems equipped with vocal interfaces (e.g., advanced cockpits, C3 Systems—Command, Control, and Communication systems, and information warfare).

The IST-TG01 (Formerly RSG.10) recognized the need to perform research and conduct studies on this topic to better understand, detect, and mitigate the effects of stress on speech production. Thereby identifying and supporting future military requirements. In order to address the different scientific aspects of this topic, Project 4 began with the organization, in cooperation with ESCA (European Speech Communication Association), of an international workshop on "Speech Under Stress" held in Lisbon, Portugal in September 1995. This very successful workshop underlined the necessity of a coordinated international effort to support NATO interests in this area. A primary outcome of this project is to create speech systems which are robust to stress indicative of the military speech environment.

In order to share the most recent advances in this field the NATO IST-TG01 established a Speech Under Stress webpage[1]. This page contains an overview of the on-going activities, collected/available speech databases, international research groups, and an extensive updated set of references. This report is organized into five chapters, with conclusions and recommendations in Chapter 7. Below we briefly summarize the main issues discussed and evaluations performed in each chapter.

Chapter 2: This chapter considers the issue of defining the problem of speech under stress. Specifically, several definitions of stress are presented with a flow diagram which connects the stage of speech production with the stress taxonomy order.

Chapter 3: This chapter presents the various speech under stress databases which were collected from participating laboratories. These databases include: SUSC-0, SUSC-1, SUSAS, DLP, vocal emotion, and DCIEM. An overall summary of the task, amount of data, language, stress type, and other characteristics is included in Table 3.2.

Chapter 4: This chapter is focussed on the analysis of speech under stress. Here, previous work and the analysis of various speech production domains are considered for the speech data discussed in Chapter 3. This chapter briefly reviews work in the field, and presents a representative set of results from the SUSAS stressed speech database.

[1]This page is located at the following Web location: `http://cslu.colorado.edu/rspl/stress.html`

Chapter 5: This chapter considers the problems of detection, classification, and assessment of input speech under stress. In many military situations, it is useful to be able to assess the stress state of the speaker. This area is also of interest to law enforcement.

Chapter 6: Here, the problems of speech recognition, speaker recognition, and speech coding/synthesis are considered with respect to speech under stress. Extensive evaluations were performed to assess the performance of commercial off-the-shelf (COTS) speech recognizers in order to illustrate the significant impact stress has on recognition performance. The tests were performed at several laboratories using speech data from SUSAS and DLP. Next, in order to address the changes in speech production under stress, several stress compensation methods are discussed which are shown to improve speech recognition performance. While these approaches have been effective in speaker dependent speech recognition applications for limited vocabularies, there has not been much success in addressing stress for COTS large vocabulary speaker independent systems. This chapter also considers the related problem of speaker recognition under stress, with experiments presented using SUSAS data. Finally, the chapter concludes with a discussion of methods and experiments of stressed speech synthesis and coding.

Chapter 7: Finally, in this chapter we draw conclusions and discuss the impact speech under stress has in military speech technology.

Chapter 2

Definitions of Speech Under Stress

Speech production is a complex process, beginning with an intention to communicate, and passing through various levels of mental processing which translate an idea into sequences of motor neuron firings, that in turn activate muscles that generate and control acoustic signals. The outcome of this process may be perturbed by so many factors that it is probably rare for two utterances to be so similar that no difference could be detected. Much (perhaps all) of the variability of speech carries information about the state of the speaker; this is generally useful in social interactions between humans but is not "understood" by machines. Human-machine interaction via speech is at present largely limited to an exchange of "words" with precisely defined meanings, and the para-linguistic content is instead a problem to be overcome. This is particularly the case when the speech is perturbed by "stress" on the talker. The aim of this chapter is to describe and explain the working definitions of stress which formed the basis of the data collection and experiments reported in later chapters. It is based on the discussions at the Lisbon workshop, as summarised by Murray, Baber and South, (1996 [119]), but some changes have been made, especially in the definitions of "second-order" and "third-order" stressors. The main focus of this work is on the effects of stress on speech production by humans, but an understanding of these effects is also useful in synthesizing speech when it is desired to simulate emotions. As an example, synthetic emotion may be added to cockpit voice warnings in order to impart a sense of urgency to the pilot. Defining "stress" is a notoriously difficult problem. In all probability, no single definition will satisfy all circumstances, or, if it does, it will be too vague to have any practical use. The definitions offered here will, we hope, be appropriate to military applications of speech technology but may be unsuited to other areas. A definition of "stress" must also be considered in the context of a model of the human system; indeed, this may be the only way to make sense of the subject. The next section considers definitions of stress in fairly general terms, while a model of speech production under stress is also described. The final section of this chapter, considers a taxonomy of stressors, as related to the speech production model.

2.1 Definitions of Stress

The title of the Lisbon workshop "Speech under Stress" was deliberately chosen to reduce confusion with the meaning of "stress" as used in the science of linguistics, i.e., emphasis given to a syllable. The implication of being "under stress" is that some form of pressure is applied to the speaker, resulting in a perturbation of the speech production process, and hence of the acoustic signal. It is often the case that the pressure is in some sense threatening to the speaker (especially in the context of military operations), but this is not always so. This definition necessarily implies that a "stress free" state exists, i.e., when all pressure is absent, although it may be hard to find circumstances completely free of stress. An important consideration of the application of speech technology is that the speech being processed is usually compared with

reference samples that were collected under unstressed conditions. This is particularly the case in military applications, where the reference speech may be recorded at home base in comfortable surroundings but during operations the users of such equipment will be subjected to noise, physical forces, fear, fatigue, etc. For practical purposes therefore, the unstressed state may be defined as that in which the reference samples of speech are collected. The term "stressor" is used to denote a stimulus that tends to produce a stress response; the actual responses produced by individuals may vary a great deal and it may not be possible to classify all possible stimuli as either stressful or not. In the context of military operations, the stressors are often easy to identify and define, and nearly always threatening to the individual's comfort and well-being.

2.2 Model of Speech Production under Stress

Speech production begins with abstract mental processes: the desire to communicate and the idea which is to be communicated. Suitable linguistic units (words or phrases) have to be chosen from memory and formed into a sentence, subject to grammatical constraints. From the abstract sequence of words, a corresponding sequence of articulatory targets must be generated, then appropriate motor programmes for the targets must be activated, with modifications to take account of context and para-linguistic information. This results in patterns of nerve impulses being transmitted to the muscles which control the respiratory system and vocal tract. The final stages are purely physical: the generation of acoustic energy, the shaping of its spectral and temporal characteristics, and its radiation from the mouth and/or nostrils. These processes are summarized on the left side of Fig. 2.1.

Although described as a sequence, these processes overlap to a considerable degree, especially in the information processing stages. It is probably quite normal for the acoustic signal to be started before the sentence has been fully formed in the higher levels of the brain. There are also many layers of feedback within the overall process, which may cause the process to be halted or re-directed at any stage if an error is found. Evidence for this is provided by the dysfluencies in normal speech (Laver [90]).

A given stressor can be considered to act primarily on a particular stage of the speech production process, and this should define its effects, within certain (possibly very wide) limits. There follows a classification of stressors based on the level in the above model at which the stressor acts (as shown on the right side of Fig. 2.1).

The stressors whose effects are easiest to understand are those which have a direct physical effect on the speech production apparatus. Examples of such "zero-order" stressors include vibration and acceleration. To a first approximation, the patterns of impulses in the motor neurons and the resulting muscle tensions will be the same as in the unstressed condition, but the responses of the articulators will change because of the external forces applied to them. There may be some modification of the neuromuscular commands as a result of auditory and proprioceptive feedback. In general, these stressors will have similar effects on all speakers, but differences in physique will affect the response, particularly under vibration when resonances in the body can magnify the effects at particular frequencies.

"First-order" stressors result in physiological changes to the speech production apparatus, altering the transduction of neuromuscular commands into movement of the articulators. These may be considered largely chemical effects whether the chemical mediators originate externally as medical or narcotic drugs, or internally as a result of illness, fatigue, dehydration, etc. Factors affecting the feedback of articulator positions would also fall into this class. Differences in individual responses could be large, especially where habituation or training is involved.

"Second-order" stressors are those which affect the conversion of the linguistic programme into neuromuscular commands. This level could perhaps be described as "perceptual" as it involves the perception of a need to change the articulatory targets, but without involving higher level emotions. The most common example is the Lombard effect, in which the stressor

is noise and the response is to increase vocal effort. This is a particular case of perception of a problem with the communication channel; other similar problems may be perceived aurally or by feedback from the listener.

"Third-order" stressors have their effects at the highest levels of the speech production system. An external stimulus is subject to mental interpretation and evaluation, possibly as a threat (as implied by the word "stress"), but other emotional states such as happiness will also have their effect at this level. Some third-order stressors may not be external stimuli at all, but originate from within the mind. These complex mental processes may affect the original idea and the construction of the sentence, which is perhaps outside the scope of this project. There will certainly be effects on the articulatory targets and neuromuscular programme expressing the emotion via paralinguistics, and possibly, through changes to physiological arousal, also at the transduction level.

The classification of stressors described above is based on the level of the speech production process at which the stressor has its primary effect. It should be borne in mind that there may also be secondary effects at other levels. For example, to someone who has never flown in a helicopter before, the high vibration levels may be perceived as a threat and result in fear, with possible third order effects from a stressor normally classified as zero-order.

2.3 Taxonomy of Stressors

This section attempts to classify various stressful stimuli within the scheme outlined above, according to their primary effects. In some cases it is not clear at which level the primary effects occur, or major effects may occur at more than one level or at different levels in different individuals. Self-awareness also makes it possible that almost all stressors may have a third-order effect. For these reasons, a definitive classification of stressors is not possible and the tentative classification offered here (Table 2.1) would need to be re-considered in the light of a particular application.

Stressor order	Description	Stressors
0	Physical	Vibration Acceleration (G-force) Personal equipment, Pressure Breathing, Breathing gas mixture
1	Physiological	Medicines, Narcotics, Alcohol nicotine, Fatigue, Sleep deprivation Dehydration, Illness, Local anaesthetic
2	Perceptual	Noise (Lombard effect), Poor communication channel, Listener has poor grasp of the language
3	Psychological	Emotion, Workload Task-related anxiety Background anxiety

Table 2.1: Stress Taxonomy

Of the physical stressors, vibration and acceleration are self-explanatory, and common in military environments. Personal equipment includes clothing and other items worn on the body, which may exert pressure on the vocal apparatus. Examples are the oxygen mask worn by fast-jet aircrew which applies pressure to the face and restricts jaw movement, or a safety harness restricting chest movement. Positive pressure breathing (used to maintain consciousness at high G-levels or in the event of loss of cabin pressure at high altitude) has the effect of distending the vocal tract and thus changing its resonant frequencies. Changes in the constituents of the breathing gas will also affect the speech signal, as when oxygen/helium mixtures are used by

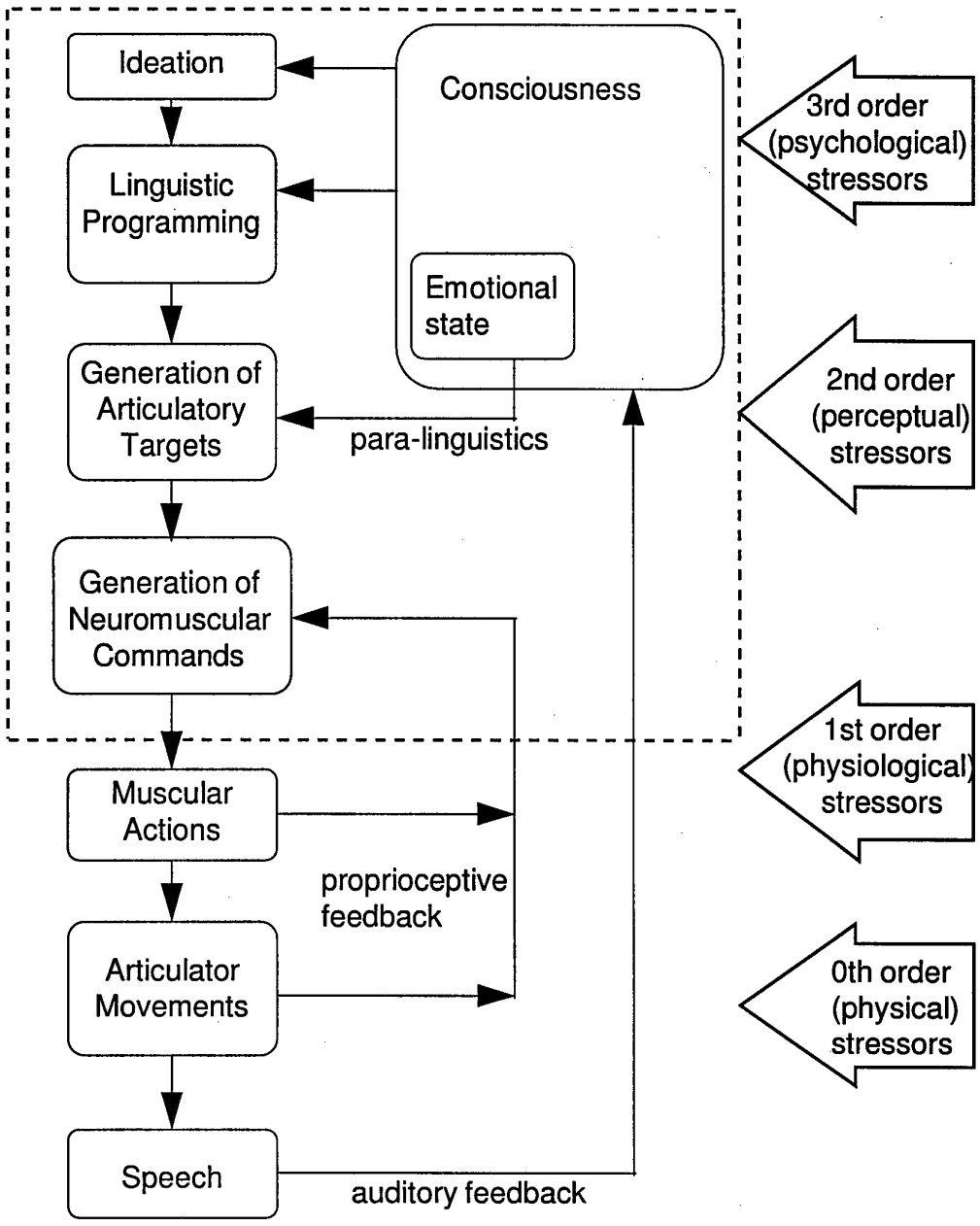


Figure 2.1: Outline of Speech Production Process and Order of Stressors.

deep-sea divers.

Physiological stressors include a wide range of stimuli, from narcotic drugs to sleep deprivation, and many of these may also have second and third order effects. Perceptual stressors form a much more limited set, but again may also have third order effects arising from, for example, frustration at the difficulty of communicating via a poor channel. The range of stimuli that could be considered psychological stressors is almost unlimited, as the input is interpreted in the light of the individual's beliefs. If we are considering military applications, speakers will in general be highly trained and mentally robust, but fear and anxiety will still be factors in most operations. Workload will probably be the most common third-order stressor, as military personnel spend the larger part of their time in training.

Chapter 3

Speech Under Stress Databases

This chapter provides a brief introduction to the speech databases used in the experiments described later in the report. More detailed descriptions can be obtained from the references or the documentation files on the CD-ROMs. In addition, sample audio demonstration files can be found on the NATO Stress Web page `http://cslu.colorado.edu/rspl/stress.html` which also includes access to additional documentation.[1]

This chapter is organized as follows. Sec. 3.1 describes the SUSC-0 database which includes stressed speech from air traffic controllers and emergency fighter cockpit environments. Sec. 3.2 considers a database of speech produced during a physical load task (climbing stairs). Next, Sec. 3.3 considers the Speech Under Simulated and Actual Stress database, which includes speech from various speaking styles, computer workload response tasks, speech produced during roller-coaster amusement park rides, and helicopter cockpit environments. Sec. 3.4 considers stress from operators reading license plates from video systems at different speeds. Sec. 3.5 considers a small database of speech created for the study of emotion for text-to-speech synthesis systems. Finally, Sec. 3.6 describes recordings made during a sleep-deprivation experiment. The main features of all the databases are summarized in Table 3.2 at the end of this chapter.

3.1 SUSC-0

This database was created specifically for this project, by bringing together existing recordings from a number of sources.

3.1.1 SUSC-0: Fighter Controllers (ground-to-air)

A study on workload and stress of fighter controllers at the Control and Report Centre of the Military Air Traffic Control Centre in the Netherlands was conducted by the TNO Human Factors Research Institute. For studies on the speech signal, communications from ground-to-air operations were recorded from 11 different speakers. For each speaker a total of 15 minutes of speech is available.

A part of this speech represents communications with fighter pilots. These communications conform to pre-defined procedures in English, but spoken generally by non-native speakers. The vocabulary is limited by the procedures. Another part of the communications are local (with other fighter controllers). These communications are normally in Dutch. The speech utterances are orthographically transcribed. The advantage of this database is that an objective measure of workload was obtained, and that physiological stress measures were recorded at 5 minute

[1]The NATO Speech Under Stress Web page was originally established by the Robust Speech Processing Lab was at Duke University at the following web location: `http://www.ee.duke.edu/Research/Speech/stress.html` RSPL has since moved to the University of Colorado, Boulder as part of a new Center for Spoken Language Understanding. The web pages continues to be maintained by RSPL (links from the old web site to the new location will also be maintained) at the new link: `http://cslu.colorado.edu/rspl/stress.html`

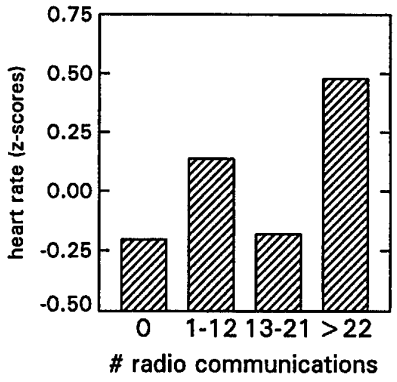


Figure 3.1: Relation between the heart rate and the number of radio contacts per 5 min. interval [49].

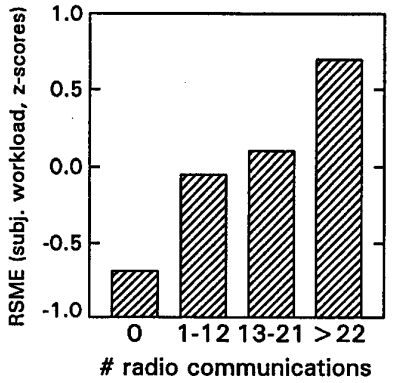


Figure 3.2: Relation between RSME (a subjective measure of workload) and the number of radio contacts per 5 min. interval [49].

intervals. In this way a 'calibrated' database can be produced. Every five minutes the following four physiological measures were derived:

- heart rate,
- blood pressure (systolic and diastolic),
- respiration (volume and rate),
- concentration of carbon dioxide in the blood (transcutaneous = $p(CO_2)$).

Additionally, the workload of the fighter controller was estimated by counting by the number of radio contacts in each five minute interval. In Fig. 3.1 the relation between the number of radio contacts and the heart rate is given. The heart rate is expressed by the z-score in order to normalize for the different rate ranges of the different speakers.

In Fig. 3.2 a similar relation between the RSME (Rating Scale Mental Effort [49]), and the number of radio contacts is given. The RSME is a subjective measure of workload as assessed by the fighter controller (also for each 5 min. interval). The graphs show a good relation between the increase of the number of radio contacts and the blood pressure and RSME respectively. This is in agreement with the relations found by Gaillard and Wientjes [49].

The results of the objective and subjective measurements are annotated to each speech utterance (typically a sentence). This type of calibrated database allows research on the relation between both objective and subjective physiological measures and speech parameters.

3.1.2 SUSC-0: Aircraft Crash

The second part of this database contains speech from communication between a fighter pilot and co-pilot which was, by accident, obtained for a stressful condition. The aircraft lost communication with the ground station due to a failure of the radio transmission switch (the transmitter was not switched off). As the aircraft was flying in ground fog (low level), it was not possible to find the runway and to land the aircraft in a safe manner. During this flight all the communications were transmitted. Voice communication starts with a conversation that the radio does not work and ends with the decision to eject from the aircraft (both pilots landed safely by parachute). Due to the use of oxygen masks by the pilots and the limitations of the radio transmission, the quality of the speech signal of this database is very poor, and hum introduced by the ground station also deteriorated the signal. The recording lasts 23 minutes, including speech and silent periods.

3.1.3 SUSC-0: F-16 Engine Out

Recording of an incident in which an USAF F-16 fighter lost engine oil pressure. Voices of the pilot of the disabled aircraft (call-sign Viper 03), another F-16 pilot (Viper 04), an air traffic controller, and some of the cockpit voice warnings are heard on the single channel recording. Viper 03 is guided towards an airfield, with Viper 04 in support, but the engine seizes up while they are still over five miles from the airfield. Fortunately, he has enough altitude to glide to the airfield and Viper 03 makes a successful engine-out landing. The quality of the recording is fair, although voices often overlap and the cockpit audio system AGC allows the background noise to come up to high levels when the pilot is not speaking. The pilot of Viper 03 sounds fairly calm at the beginning, becomes very excited when his engine seizes, and is clearly very relieved after landing. The recording is 15 minutes long.

3.2 SUSC-1

3.2.1 SUSC-1: Physical Stress Database

For analysis on the effect of physical stress on speech production, subjects were asked to pronounce a short sentence twice after a fair physical load (running up and down a staircase of three floors for ten times). Just before and just after this exertion, recordings were made in a soundproof room with a high quality recording system.

For the experiments 10 male and 10 female speakers were used. Each subject performed the task twice. This concept was designed to compare the effect on the speech production of physical stress versus "relaxed" speaking, gender, and inter speaker variations. In addition to this, for two speakers the recordings were repeated 10 times on 10 different days. This gives some data to study the effects of intra speaker variability. All the speech material was annotated at phone level.

3.3 SUSAS: Speech Under Simulated & Actual Stress Database

This database was established in order to conduct research into the analysis and recognition of speech produced in noise and under stress (Hansen [54]). *SUSAS* refers to "Speech Under Simulated and Actual Stress," and has been employed extensively in the study of how speech production and recognition varies when speaking during stressed conditions (see references [60, 68, 58, 54]). SUSAS consists of five domains, encompassing a wide variety of stresses and emotions (see Fig. 3.3). A total of 32 speakers (13 female, 19 male) were employed to generate in excess of 16,000 isolated-word utterances. The five stress domains included were: (i) talking

SUSAS DATABASE

SPEECH UNDER SIMULATED AND ACTUAL STRESS

DOMAIN	TYPE OF STRESS OR EMOTION	SPEAKERS	COUNT	VOCABULARY
TALKING STYLES	SIMULATED STRESS SLOW SOFT FAST LOUD ANGRY CLEAR QUESTION	9 SPEAKERS (ALL MALE)	8820	35 AIRCRAFT COMMUNICATION WORDS
SINGLE TRACKING TASK	CALIBRATED WORKLOAD TRACKING TASK: MODERATE & HIGH STRESS LOMBARD EFFECT	9 SPEAKERS (ALL MALE)	1890	35 AIRCRAFT COMMUNICATION WORDS
DUAL TRACKING TASK	ACQUISITION & COMPENSATORY TRACKING TASK: MODERATE & HIGH STRESS	8 SPEAKERS (4 MALE) (4 FEMALE)	2257	35 AIRCRAFT COMMUNICATION WORDS
ACTUAL SPEECH UNDER STRESS	AMUSEMENT PARK ROLLER-COASTER HELICOPTER COCKPIT RECORDINGS (G-FORCE, LOMBARD EFFECT, NOISE, FEAR, ANXIETY)	11 SPEAKERS (4 MALE, 3 FEMALE) (4 MALE)	1642	35 AIRCRAFT COMMUNICATION WORDS
PSYCHIATRIC ANALYSIS	PATIENT INTERVIEWS: (DEPRESSION, FEAR, ANXIETY, ANGRY)	8 SPEAKERS (6 FEMALE) (2 MALE)	600	CONVERSATIONAL SPEECH: PHRASES & SENTENCES

Figure 3.3: SUSAS: Speech Under Simulated and Actual Stress database.

styles[2] (slow, fast, soft, loud, angry, clear, question), (ii) single tracking computer response task or speech produced in noise (Lombard effect), (iii) dual tracking computer response task, (iv) subject motion-fear tasks (G-force, Lombard effect, noise, fear), and (v) psychiatric analysis data (speech under depression, fear, anxiety). A common highly confusable vocabulary set of 35 aircraft communication words make up the database (subsets include {go, hello, oh, no}, {six, fix}, etc. See Table 3.1). Each subsection of the database is described briefly below; a more complete discussion of SUSAS can be found in the literature (Hansen [60, 58, 54]; Hansen and Bria [66]; Hansen and Cairns [68]).

All speech tokens were sampled using a 16-bit A/D converter at a sample rate of 8 kHz. All speech files have been orthographically transcribed and labeled at the phone and word levels using a new parsing routine [125, 126]. Further details are contained in the following document: "Getting Started with the SUSAS: Speech Under Simulated and Actual Stress Database," J. Hansen, S. Bou-Ghazale, R. Sarikaya, and B. Pellom, Robust Speech Processing Laboratory, Technical Report: RSPL-98-10, April 15, 1998 (contained on the CD-ROM) and in the reference [65][3]. We briefly summarize the stress areas below.

3.3.1 SUSAS: Talking Styles Domain

The first SUSAS domain involves speech under various speaking styles. Data in this domain was originally used in studies by Lippmann, et al. [95, 124] (16 kHz sampled), and was donated by R.

[2]We note here that approximately half of this portion of SUSAS consists of style data donated by Lincoln Laboratory (Lippmann, et al. [96]).

[3]The SUSAS Stressed Speech Database from RSPL is available from the Linguistics Data Consortium at the following web location: `http://morph.ldc.upenn.edu/Catalog/LDC99S78.html`

35-Word SUSAS VOCABULARY SET					
brake	eighty	go	nav	six	thirty
change	enter	hello	no	south	three
degree	fifty	help	oh	stand	white
destination	fix	histogram	on	steer	wide
east	freeze	hot	out	strafe	zero
eight	gain	mark	point	ten	

Table 3.1: A summary of the 35-word vocabulary set used for SUSAS domains 1 to 4 (talking styles, single tracking task, dual tracking task, actual speech under stress).

Lippmann and C. Weinstein of Lincoln Laboratory. This portion of SUSAS contains utterances under eight speaking styles (normal, slow, fast, soft, loud, question, clear enunciation, angry). The words were produced by nine male talkers sampling three major accents (General American, Boston, New York). Each word was produced 28 times by each subject. Both training and test data is included.

3.3.2 SUSAS: Single Tracking Task Domain

The second SUSAS domain consists of speech data from a single tracking task, and speech under Lombard effect. The same 35-word vocabulary and 9 speakers were used as in the talking styles domain. Speech was produced while performing a computer workload task which stimulates stress, originally proposed by Jex [81]. In this approach, a single tracking task, which reflects graded levels of mental workload, was developed. The operator views a display of the error between the command input and plant output, which is the response of a marginally stable, single-pole linear system, and corrects errors with opposite pressure on a control stick (similar to a joy-stick in some video games). The degree of instability may be adjusted for varying degrees of difficulty. Two levels of workload difficulty were used in this section. Subjective ratings, performance data, and heart rate data indicated that the high workload ($\lambda = 70\,\%$) was measurably more difficult than the moderate level ($\lambda = 50\,\%$). This domain also includes a small portion of speech produced in noise, simulating the Lombard effect [100]. The Lombard effect occurs when talkers vary their speech characteristics in order to increase intelligibility when speaking in a noisy environment. For this portion, Pink noise was presented binaurally at an overall level of 85 dB SPL.

3.3.3 SUSAS: Dual Tracking Task Domain

The third SUSAS domain consists of speech produced while performing a dual tracking task (simulating flight control and target acquisition) as a means of inducing workload. This task addresses a pilot's two key goals and was developed by Folds, et al. [45, 46] for the USAF School of Aerospace Medicine. Task difficulty could be controlled by time constraints for completion, or increasing resource competition or motivation. The primary tracking task was a pursuit task in which the input signal was determined by the sum of two sine functions. The response marker (output signal) was a small circle. The vertical position of the circle was fixed at the center of the display; the horizontal position was determined by the movement of a control stick in its x-axis. The trail of response markers (triangles and circles) indicates past attempts by the operator to perform the task. Speech data was collected during the performance of the dual-tracking tasks at two different levels of workload stress.

3.3.4 SUSAS: Actual Speech Under Stress

The fourth domain consists of speech produced during the completion of two types of subject motion-fear tasks. These tasks were chosen because they required no training, yet generate the type of stress (fear or anxiety) which might be experienced in an aircraft cockpit emergency

situation. Two rides from an amusement park were chosen: the *Scream Machine* and *Free Fall.* The *Free Fall* ride lasts for about 60 seconds, with the free fall portion comprising about 10 seconds. Four seated passengers are strapped in an upright seated position into a car which is raised vertically to approximately 130 feet. The car drops vertically downward for about 100 feet, before rolling onto a horizontal portion of the track for deceleration. During the free fall portion, talkers repeated several pre-chosen words. The second task was the *Scream Machine*, which is a typical wooden frame roller-coaster which seats roughly 30–36 passengers. Due to the large number of passengers, higher levels of background screaming can be heard in these recordings. The overall ride consists of large vertical movements with small amounts of lateral movement during calm periods between drops. The entire ride lasts for about 90 seconds. Each speaker performed the task twice. The speaker's location during the ride was identified based on timing and background noise. Each recording was partitioned and subjectively marked for stress with respect to time and position during the task. The chosen subjects were all native Americans with no apparent speech deficiencies. A total of 1642 utterances were collected from speech under task stressed and baseline neutral recordings. In each subject motion task, at least four factors contributed to the type of speech recorded: g-force variation, background noise, Lombard effect, fear and/or anxiety.

While SUSAS was organized and collected during the period 1985–88, (Hansen, 1988 [54]), it was determined that additional speech under actual stress would be useful. Four additional male speakers were later added in 1993 using the same 35-word vocabulary which included pilots flying missions in Apache helicopters. Two of the pilots were operating their Apache helicopter in normal flight conditions as (i) baseline with helicopter on the ground but running, and (ii) pilots flying their helicopter while speaking. An additional set of recordings were included from two other Apache helicopter pilots flying a night mission into the Raleigh/Durham Airport while running low on fuel (the pilots were not familiar with the area, since they were from another state in the U.S.). This speech consists of tactical communications between pilot, co-pilot and sometimes an air-control person giving directions. The speech data consists of continuous speech passages not from the 35-word vocabulary. Stress levels increase as the fuel level begins to drop. Two large digitized files are included which contain speech from each pilot and co-pilot.

3.3.5 SUSAS: Psychiatric Analysis Domain

The last SUSAS domain is in psychiatric analysis. A collection of recordings were obtained from a Emory Medical Center, Department of Psychiatry for the purposes of obtaining examples of speech under emotional stress. Patients undergoing psychiatric analysis were recorded using a high quality microphone and tape recorder in a natural doctor-patient environment. This data was not released since it contains direct patient information, however it is available upon special permission (Hansen [54, 60]).

3.4 DLP: License Plate Reading Task

The entry of alpha-numeric data is highly relevant to many military tasks, from entering NATO stock numbers in the storeroom to aircraft identification in the field. The majority of the speech databases used for recognition system evaluation are obtained in highly controlled conditions to reduce variability. These databases generally make use of scripted speech and exclude large variations in speech style and spoken errors. A more representative corpus would involve the recording of subjects carrying out a realistic data entry task, which will include, rather than exclude, spoken errors and speech level and rate variations.

There was thus a requirement to record a speech corpus of alpha-numerics spoken under a suitable simulation of a 'real' data entry task. It was desirable to ensure that the subjects concentrated more on the data entry task than providing a good speech performance. An alpha-

numeric entry task with a potential relevance to required military functions was considered to be the reporting of car number plates.

The DERA License Plate (DLP) corpus was recorded at the Speech Research Unit during the summer of 1992. It comprises recordings of car number plates read using the phonetic alphabet of the International Civil Aviation Organisation (ICAO) (alpha, bravo, charlie etc.), plus digits. Fifteen speakers were recorded, 11 male and four female, covering a range of ages and featuring a number of non-extreme British regional accents. The number plates were presented via video playback through a monitor, with two different rates of presentation, fast and slow. The fast rate was intended to be above the dictation rate of the subjects and hence a source of cognitive stress. The subjects' perceived stress during the dictation sessions was recorded to provide data for a corpus of speech under limited stress conditions.

This task was chosen partially to fulfill the demand for an initial speaker independent test database of spoken alpha-numerics, and partly to examine the effects of cognitive stress on speech performance and any concurrent effects on automatic speech recognition performance.

The database was recorded in conjunction with the Industrial Ergonomics Group at Birmingham University, whose interest was in examining dictation accuracy, speaker coping strategies and the effects of stress on the performance of speech recognizers

3.4.1 DLP Database: Details of The Task

3.4.1.1 Experimental Task

The task chosen for the recordings was the dictation of British car number plates. The data was collected by recording vehicles' number plates onto video, with the camera panning across the fronts of the cars at a constant speed. The video was then played back to the speakers for dictation. The data was recorded in a car park, with the car number plates clearly visible and uniformly space separated. Each number plate set covered ninety seconds of data. There were a total of 159 number plates across all the recordings, with no repetition of number plates.

3.4.1.2 Dictation Task Details

The rate of viewing of the number plates was, on average, 787 number plates per hour for the slow rate and 1309 for the fast. Three sets of recordings were made for each of the two speed conditions. The database thus comprises 90 speech files totaling 135 minutes of speech from 15 speakers. Each subject was shown the clips in a randomized order. After each individual test, the subject was asked to complete a NASA-TLX subjective workload questionnaire to provide a subjective measure of the stress.

Display of the video data was via a large monitor placed for easy viewing and such that the car number plates were clearly visible. The subjects were required only to dictate the car number plates as accurately as possible. No other tasks were required. No feedback with regards to their vocal performance was provided.

The number plate data was shown un-rehearsed and without repetition to ensure that no familiarity was possible. No attempt was made to correct spoken errors; all spoken errors are included in the database.

3.4.1.3 Speaker Experience

The speaker population comprised eleven male and four female speakers aged 20 to 38 years, plus an additional trial recording by a male. Accents were all non-extreme British English from a variety of regions. All the speakers were working on speech or signal processing topics either full-time or as short-term summer studentships. All subjects were made aware of the aims of the recordings in advance.

The translation of number plates to the ICAO form requires a level of cognitive processing, such that subjects with varying familiarity with the vocabulary will experience different levels of stress for the same conditions. The ability of the subjects was such that with no time constraints, they could dictate given number plates without translation errors. At the fast rate the subjects were not able to dictate all the number plates, causing an increase in errors and stress.

3.4.1.4 Data Processing

Recording was carried out in a sound proof booth, with the speech signal recorded onto DAT with a sampling rate of 48 Khz from a SHURE SM10A headset mounted microphone. Recording level was equalized across all speakers by varying the recording gain before each session on representative speech. Background noise was low (SNR better than 42 dB) across most of the speakers, though some files were recorded with a low level of 50 Hz hum evident.

3.4.1.5 Annotation of the DLP Corpus

The speech data was down-sampled from DAT to computer hard disk at 20 kHz and archived to optical disk as signal format files. The speech files were filterbank analyzed using the standard 27 channel SRUBANK analyzer.

A detailed orthographic transcription was made of the speech data which included all the out-of-task speech. Annotation of the filterbank domain files was carried out by forcing recognition of the speech files against the orthographic transcription, using the ASTREC speech recognizer. The speech models used for the recognition were taken from the SI89 database of air reconnaissance task recordings and comprised a speaker independent model set of alpha-numeric tokens. The annotations were then hand checked to verify the automatic process.

3.5 Databases for the Study of Vocal Emotion in Speech Synthesis

Most studies of emotions are based either on a "palette" model, a multi-dimensional model or combinations of both. The palette model defines a closed set of "basic" emotions from which all other emotions may be derived; the multi-dimensional model defines emotions as points within some form of dimensional space. Although the number and type of basic emotions may differ substantially from one author to another, a typical palette may include: anger, happiness, sadness, fear and disgust. Hence, a few databases designed for the training and evaluation of speech-with-emotion synthesis systems include these sentences with primary emotions. A database designed by Murray (1995 [118]) was designed for the study of 6 emotions (anger, happiness, sadness, fear, disgust and grief). It includes 39 different phrases which were spoken in different ways by a synthetic speech system and used in perceptual experiments to evaluate the capability of synthesizing emotions of this system. The 39 phrases were subdivided into:

- 18 neutral phrases (3 for each of the 6 emotions)
- 21 emotionally loaded phrases (3 for each of the 6 emotions, plus neutral)

The perceptual experiments involved:

- 18 neutral phrases spoken in a neutral voice
- 21 emotionally loaded phrases spoken in a neutral voice
- the same 18 neutral phrases spoken with one of the 6 emotions
- the same 21 emotionally loaded phrases spoken with the appropriate vocal emotion

Other speech databases have also been formulated to study emotion for speech synthesis. An emotional speech database for Spanish has been recorded at Universidad Politecnica de Madrid (UPM) using one actor, to study anger, happiness, sadness and surprise. The database contains

3 types of structures: short neutral sentences, single words or short phrases extracted from the previous group and paragraphs in which some of the short sentences may be embedded.

Finally, the BDFALA database [105] was collected for European Portuguese which includes 29 emotionally-loaded sentences, each read by 10 speakers[4].

The sentences were designed to study anger, happiness, irony and disgust. A new database is currently being recorded with both neutral and emotionally-loaded sentences for 5 basic emotions.

3.6 DCIEM Sleep Deprivation

This database was collected during the DCIEM Sleep Deprivation Study (1994). The Defense and Civil Institute of Environmental Medicine (Department of National Defense, Canada) has a history of studies into the effects of sleep deprivation. This database was directed towards effects of drugs on performance decrements during sleep deprivation. It is known that amphetamines counter some decrements, but amphetamines have undesirable side-effects. For some time, the French Army had been making available to its soldiers on demand a drug called Modafinil, which is prescribed for narcolepsy, but which had not been tested for its effect on the performance of otherwise normal individuals deprived of natural sleep. Unlike amphetamines, Modafinil is known to have a wide ratio between useful and toxic doses. The present study was designed to determine whether Modafinil would work as well as amphetamines in reducing the performance decrements associated with sleep loss.

It consists of a map reading task which is based on the HCRC Map Task Corpus. The map task is a cooperative task involving two participants. The two speakers sit opposite one another and each has a map which the other cannot see. One speaker, the Instruction Giver, has a route marked on his/her map, while the other, the Instruction Follower, has no route. The speakers are told that their goal is to reproduce the Instruction Giver's route on the Instruction Follower's map. The maps are not identical and the speakers are told this explicitly at the beginning of their first session. It is, however, up to them to discover how the two maps differ. No restrictions are place on what either speaker can say. All dialogues were recorded via close-talking microphones, with one channel per speaker, on a Panasonic SV-3500 DAT recorder in quiet conditions. All participants took part in a number of sessions, and so gained experience with different maps. The maps themselves differ as a result of the systematic manipulation of the following design variables: (1.) phonological characteristics of feature names, (2.) the extent to which features contrast or are shared between the maps. The assignment of speakers to maps involves two further variables: (3.) drug condition (Modafinil, Amphetamine, and Placebo), and (4.) subgroup size. An unusual feature of this study is that subjects' time was almost wholly occupied with the performance of various psychological tasks. These were scheduled in 6-hour BLOCKS of three 2-hour SESSIONS each. The task sequence was the same in each 6-hour block, but differed among the 2-hour blocks in some respects. The first day of the study was used for introduction to the tasks. Blocks began on the second day and continued throughout, except for sleep periods on second, fifth, and sixth nights. The speech data is available from the Linguistics Data Consortium (LDC).

[4]This database is available from the authors of [105]

	SUSC-0			SUSC-1	DERA
	Fighter Cockpit & Controller	Aircraft Lost in Fog	F-16 Engine Out	Physical Exertion	Licence Plates
Stressor	Workload	Anxiety	Anxiety	Physical Exertion	Pressure
Stressor Type	Psych.	Psych.	Psych.	Physio.	Psych.
Stressor Order	3	3	3	1	3
Language	English and Dutch	Mostly Dutch, Some English	US English	Dutch	UK English
Task	Command and Control	Fighter Cockpit	Fighter Cockpit Air Traffic Control	Read Sentences	Prompted Phrases
Quantity	3 hours	23 mins	15 mins	15 mins	2 hours
Number of Speakers	9	2	3	20	16
Gender	Male	Male	Male	10 Female 10 Male	4 Female 12 Male
Native Language	Dutch	Dutch	US English	Dutch	UK English
Population	Air Force Controllers	Pilots Pilots	Pilots and ATC	Researchers	Researchers Time
Microphone		Oxygen Mask	Oxygen Mask		Shure SM-10
Sampling Rate	16 kHz	16 kHz	16 kHz	16 kHz	20 kHz
Recording Quality	Good	Poor	Poor	Good	Good

	SUSAS					
	Talking Styles	Lombard Effect	Computer Tracking Tasks	Roller-Coaster Rides	Helicopter	
					Word Commands	Fuel Low
Stressor	Emotion	Noise	Time Pressure, Workload	Acceleration Exhilaration	Noise, Vibration	Noise, Vibration Anxiety
Stressor Type	Psych. (simulated)	Perceptual	Psych.	Physical and Psych.	Physical, Psych. and Perceptual	Physical, Psych. and Perceptual
Stressor Order	3	2	3	0, 3	0, 2 and 3	0, 2 and 3
Language	US English	US English	US English	US English	US English	US English
Task	Isolated Words	Isolated Words	Isolated Words	Isolated Words	Isolated Words	Spontaneous Phrases
Quantity	2 hours	30 min	50 min	45 min	30 mins	15 min
Number of Speakers	9	9	8	7	4	2
Gender	Male	Male	4 Male 4 Female	3 Female 4 Male	Male	Male
Native Language	US English	US English	US English	US English	US English	US English
Population	Adults	Adults	Graduate Students	Graduate Students	Aircrew	Aircrew
Microphone	Sennheiser HMD-224	Sennheiser HMD-224	Shure 512	Shure 512	boom	boom
Sampling Rate	8 kHz	8 kHz	8 kHz	8 kHz	8 kHz	8 kHz
Recording Quality	Good	Good	Good	Variable (noisy)	Variable (noisy)	Poor (noisy)

Table 3.2: Summary of All Stressed Speech Databases.

Chapter 4

Analysis of Speech Under Stress

Due to difficulty in experimental design and limited research efforts, changes in the characteristics of speech produced under workload stress remain unclear. Thus far, research has been limited in scope, often using only one or two subjects and analyzing a single parameter (often fundamental frequency or pitch). It is not unusual for researchers to report conflicting results, due to differences in experimental design, level of actual or simulated stress, or interpretation of results. For example, some studies concentrated on analysis of recordings from actual stressful situations (Kuroda, et al., 1976 [89]; Simonov and Frolov, 1977 [144]; Streeter, et al., 1983 [150]; Williams and Stevens, 1972 [160]), while others used simulated stress or emotions (Hecker, et al., 1968 [75]; Hicks and Hollien, 1981 [79]; Williams and Stevens, 1972 [160]). This offers the advantage of a controlled environment, where a single emotion can be examined with little background noise. In some cases, variable task levels of stress have been used. Other advantages include larger data sets with multiple speakers. This allows results to be based on general speaker characteristics instead of possibly particular characteristics of an individual speaker in conveying emotion. The major disadvantage in these studies has been the reduction in task stress levels. In addition, studies using actors may produce exaggerated caricatures of emotions in speech.

In this Chapter, we summarize a series of studies which have considered the analysis of speech under stress. Sec. 4.1 briefly considers several background studies on speech under stress associated with acceleration and vibration. Sec. 4.2 presents an overview of a number of studies conducted on the SUSAS speech under stress database. Finally, Sec. 4.3 considers analysis of the SUSC-0 speech corpus.

4.1 Stress Effects of Noise, Acceleration, and Vibration

Noise is certainly one of the major stressors encountered in military environments. In high performance fighter aircraft levels of 115–120 dB are not uncommon. Even with the 15–20 dB attenuation provided by the flight helmet and earcups, levels of up to 100 dB can be reached at the pilot's ears. This induces what is called the Lombard effect (Lane, et al., 1970 [91]; Lane and Tranel, 1971 [92]). This effect causes speakers to increase the volume of their speech and to increase their fundamental frequency. Pisoni, et al. (1985) [128] and Bond, et al. (1986) [10] in separate experiments reported that in addition there are effects on the formants. The vowel space defined by the first two formants (F1 and F2) becomes smaller and the distribution of energy within the speech spectrum shows an increase in the high frequency third formant (F3) region (know as spectral tilt). In terms of the effects on speech recognition performance, Rajasekaran, et al., (1986) [131] has reported that the effect of noise at the speaker's ears results in a greater degradation in the performance of an automatic speech recognition system than does the presence of noise at the speaker's microphone. Stanton (1988) [146, 147] reported results comparing loud and Lombard speech. His results showed that Lombard speech caused

greater degradation in performance than did loud speech (Stanton developed a slope-dependent weighting technique that reduced the degradation error rate of both loud and Lombard speech by 50 %). The implication of these studies is that the changes in speech production attributable to a high ambient noise environment are sufficient to effect the performance of speech recognition systems.

Another environmental stressor that is experienced in high performance aircraft is acceleration. Acceleration effects have been demonstrated on respiration and motor control, both of which may influence speech production. Bond, et al. (1987) [11] examined speech produced by two male speakers under normal ($+1g$) and at high-sustained acceleration ($+6g$) to determine the effects of acceleration on the acoustic-phonetic structure of speech. Speakers wore oxygen masks and breathing was supplied through a chest-mounted regulator. Increases were found in fundamental frequency for both speakers. No increase in amplitude was found, but the vowel space as defined by F1 and F2 became more compact. Although no recognition experiments were conducted, the change in vowel space could effect performance of speech recognition systems.

The final environmental stressor of concern is that of vibration. Moore and Bond (1987) [108] reported preliminary results using a laboratory database collected for initial ground based evaluations of ASR systems to be flown on the AFTI F-16 (Anderson, et al., 1985) [4]. The speakers wore oxygen masks and chest-mounted regulators and were exposed to four experimental conditions, a control (no vibration), low, medium, and high levels of vibration. The levels emulated the buffeting that might be encountered on a low level, high speed flight. The results indicated that once again fundamental frequency increased and the vowel space becomes more compact. These effects are in addition to the modulation or "shakiness" imposed on the voice due to the whole body vibration. More extensive work by Bond and Moore (1990) [12] also showed decreases in spectral tilt (increase in the high frequency energy). Laboratory studies (Dennison, 1985 [43]; Cruise, et al., 1986 [35]) as well as data collected in flight (Malkin and Dennison, 1986 [101]) indicated that helicopter vibration environments at that time did not substantially affect the performance of the ASR systems tested. However, evaluations performed in Great Britain with speech recorded at vibration levels greater than those used in the above studies reported a decrease in performance of an ASR system. The vibration level used for this study were levels that the author considered reasonable for next generation high performance rotary wing aircraft (Leeks, 1986 [93]).

4.2 Production & Recognition Based Feature Analysis using the SUSAS Database

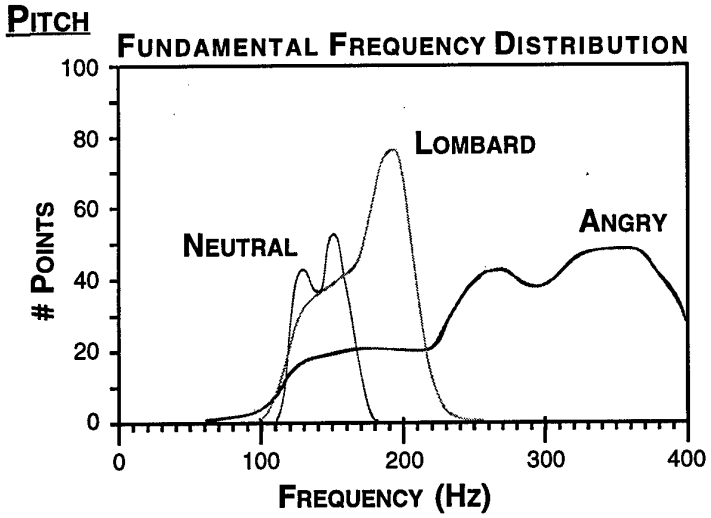
This section discusses several results from a comprehensive investigation of acoustic correlates of speech under stress using the SUSAS database (Hansen, 1988-96 [54, 55, 59, 60]). In these studies, well over 200 parameters and 10,000 statistical tests were considered in evaluating the following parameter areas of speech production: (i) pitch, (ii) duration, (iii) intensity, (iv) glottal source, (v) and vocal tract spectrum.

4.2.1 Pitch

The most widely considered area of stress evaluation are characteristics of pitch. These studies have considered subjective assessment of pitch contours, statistical analysis of pitch mean, variance, and distribution (see Fig. 4.1) [54, 60]. A partial list of conclusive points are:

- Mean pitch values may be used as significant indicators for speech in soft, fast, clear, Lombard, question, angry, or loud styles when compared to neutral conditions.
- Loud, angry, question, and Lombard mean pitch are all significantly different from all other styles considered.

- Speech produced under Lombard effect gave mean pitch values most closely associated with pitch from fast and clear conditions.
- Soft and loud pitch variance are significantly different from all styles considered.
- Pitch variance for clear and Lombard conditions are similar, but different from all other styles considered.



	NEUTRAL	LOMBARD	ANGRY
MEAN (HZ)	145	160	253
STAND. DEV.	15	24	95

Figure 4.1: Sample pitch (fundamental freq.) distribution variation for speech under neutral, Lombard, and angry stress conditions.

4.2.2 Duration

Previous studies of speech under stress have not considered statistical evaluations of individual phoneme class duration. Duration analysis was conducted across (i) whole words, and (ii) individual speech classes (vowel, consonant, semivowel, and diphthong) [54, 60]. An analysis was also conducted on interclass duration movement to determine if speakers increased duration of certain phoneme classes at the expense of others. Examples of overall word phone class (vowel, semivowel, and consonant) duration for neutral, angry, and Lombard stressed speech can be found in Figure 4.2. The length of the bars indicate the overall change in word duration. The numbers within each bar indicate the percentage of time spent in that phone class and the arrows represent a statistically significant shift in the phone class duration ratio between phone classes. A partial list of duration conclusions are:

- Mean word duration was a significant indicator for speech in slow, clear, angry, loud, Lombard, or fast styles when compared to neutral.
- Slow and fast mean word duration were all significantly different from all other styles considered.
- Clear mean consonant duration was significantly different from all styles except slow.
- Duration variance increased for all domains (word, vowel, consonant, semivowel, diphthong) under slow stress.
- Duration variance decreased for most domains under fast stress condition.
- Duration variance significantly increased for angry speech.
- Clear consonant duration variance was significantly different from all styles.

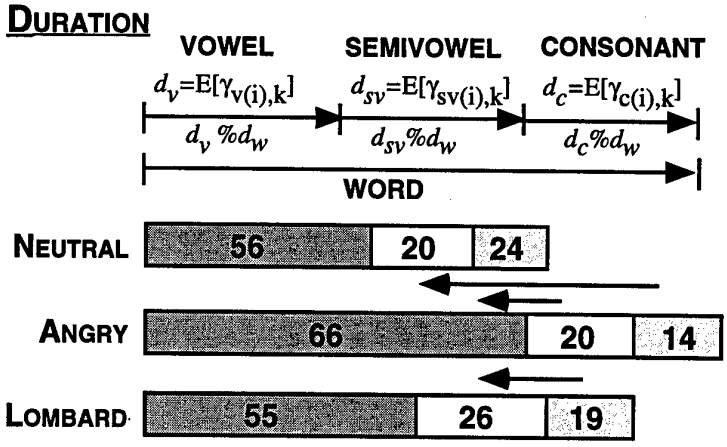


Figure 4.2: Sample duration variation for speech under neutral, Lombard, and angry stress conditions.

4.2.3 Intensity

An analysis was conducted on (i) whole word intensity, and (ii) speech phoneme-class intensity (vowel, semivowel, diphthong, consonant) [54, 60]. Statistical tests were performed on mean, variance, and distribution across the database. The shift in available energy between speech classes was also considered to determine if speakers reduce intensity in some classes in order to increase others. A partial list of conclusions are:

- Average RMS word intensity values were significant indicators for speech in angry, loud, and high workload task styles when compared to neutral conditions.
- Loud and angry RMS word intensity were significantly different from all other styles considered.
- Loud and angry RMS vowel and diphthong intensities were significantly different from all other styles considered.
- RMS consonant and semivowel intensity were not significant stress indicators for any of the styles considered.
- Variance of average RMS word intensity values were significant indicators for speech in angry and loud styles when compared to neutral.
- Variance of loud and angry average RMS word intensity was significantly different from most other styles considered.

4.2.4 Glottal Source

Aspects such as duration of each laryngeal pulse (open/closed periods), instant of glottal closure, spectral structure of each glottal pulse, or pulse shape play important roles in conveyance of stress state (Hansen, 1988; Cummings and Clements, 1989, 1990, 1995; [54, 36, 37, 39]). Due to limitations of glottal inverse filtering techniques in stress evaluation, this portion focused on direct estimation of the glottal flow spectrum. Examples of spectral structure, average spectral value, and spectral slope (in decibels/octave) are shown in Fig. 4.4. The present analysis of glottal source spectrum revealed that parameters such as spectral slope and the distribution of energy are important for relaying stress.

4.2.5 Vocal Tract Spectrum

Analysis of vocal-tract spectrum focused on formant location and bandwidth for selected phonemes across the SUSAS database. Mean and variance estimates for specific phonemes were

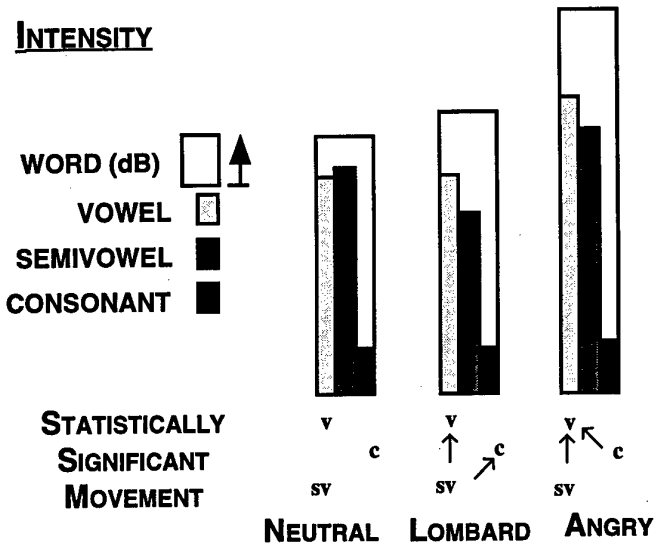
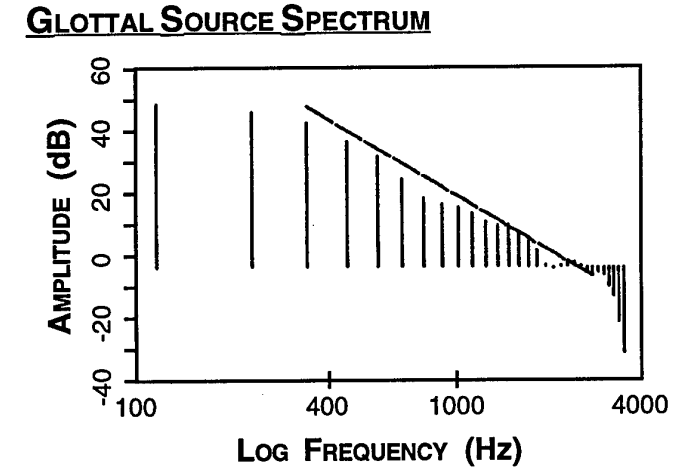


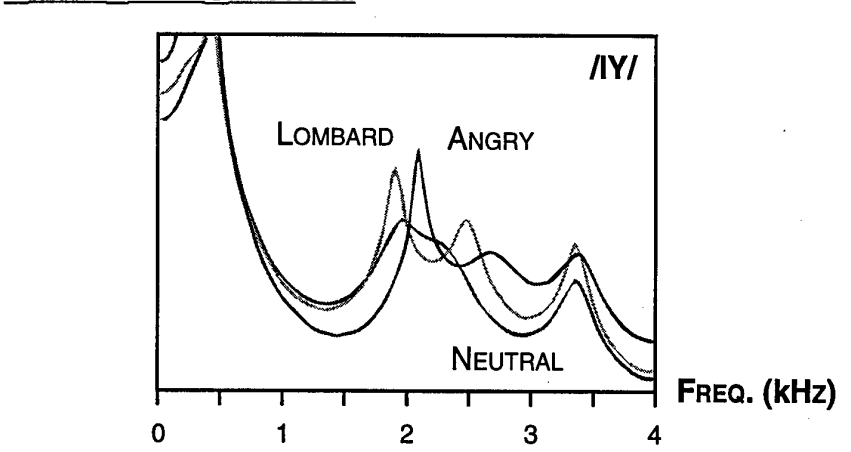
Figure 4.3: Sample RMS intensity (dB) variation for speech under neutral, Lombard, and angry stress conditions.



	NEUTRAL	LOMBARD	ANGRY
AVG. SPEC. (dB)	15.2	20.8	23.0
SLOPE (dB/OCT.)	-12.1	-9.2	-9.4

Figure 4.4: Sample glottal source spectra for speech under neutral, Lombard, and angry stress conditions.

analyzed for the 11 stress conditions. Statistical evaluations showed that of the 400 Student t-tests, 166 were statistically different from neutral. Most of these involved loud, angry, or Lombard effect formant information. A majority of the significant comparisons involved mean and variance of formant location and bandwidth for F1 and F2 as shown in Fig. 4.5. Of the ten stress conditions, average formant information for loud and angry were the most consistent across the phonemes tested.



	F1 (Bw)	F2 (Bw)	F3 (Bw)	F4 (Bw)
NEUTRAL	411 (52)	1970 (222)	2607 (496)	3368 (366)
LOMBARD	412 (73)	2006 (139)	2644 (250)	3376 (185)
ANGRY	586 (102)	2078 (166)	2661 (464)	3357 (392)

Figure 4.5: Sample vocal tract spectra variation for speech under neutral, Lombard, and angry stress conditions.

4.2.6 Vocal Tract Articulatory Profiles

The speech features considered thus far focus on characteristics of speech production. As another analysis area on speech under stress, it would be beneficial to illustrate how stress effects vocal tract structure [62]. This analysis is based upon a linear acoustic tube model with speech sampled at 8 kHz. One means of illustrating the effects of stress on speech production is to visualize the physical vocal tract shape. The movements throughout the vocal tract can be displayed by superimposing a time sequence of estimated vocal tract shapes for a chosen phoneme. Wakita, 1973 [157] proposed a method to estimate vocal tract shape using an acoustic model. The vocal tract shape analysis algorithm assumes a known normalized area function and acoustic tube length. The algorithm begins by computing the sagittal (vocal tract length) distance function by assuming a cylindrical vocal tract. Next, a set of rigid points from the glottis to the upper teeth (and rigid upper lip) models the hard palate. With the hard palate model in place, the soft palate and pharynx are approximated by forming a dependence upon the sagittal distance function. Finally, the lower lips are modeled using one of four rigid models dependent upon the acoustic tube length.

The articulatory model approach by Wakita, 1973 [157] was used to consider changes in vocal tract shape under neutral and various stress conditions as illustrated in Fig. 4.6. Here, a set of vocal tract shapes are superimposed for each frame in the analysis window. For *Neutral*, there is some movement of the articulators in the pharynx and oral cavities (as there should be for the production of the /r-iy/ phone sequence in "freeze"). There is also limited movement for the *Soft* speaking condition. However, for *Angry* and *Lombard* conditions, there is significant

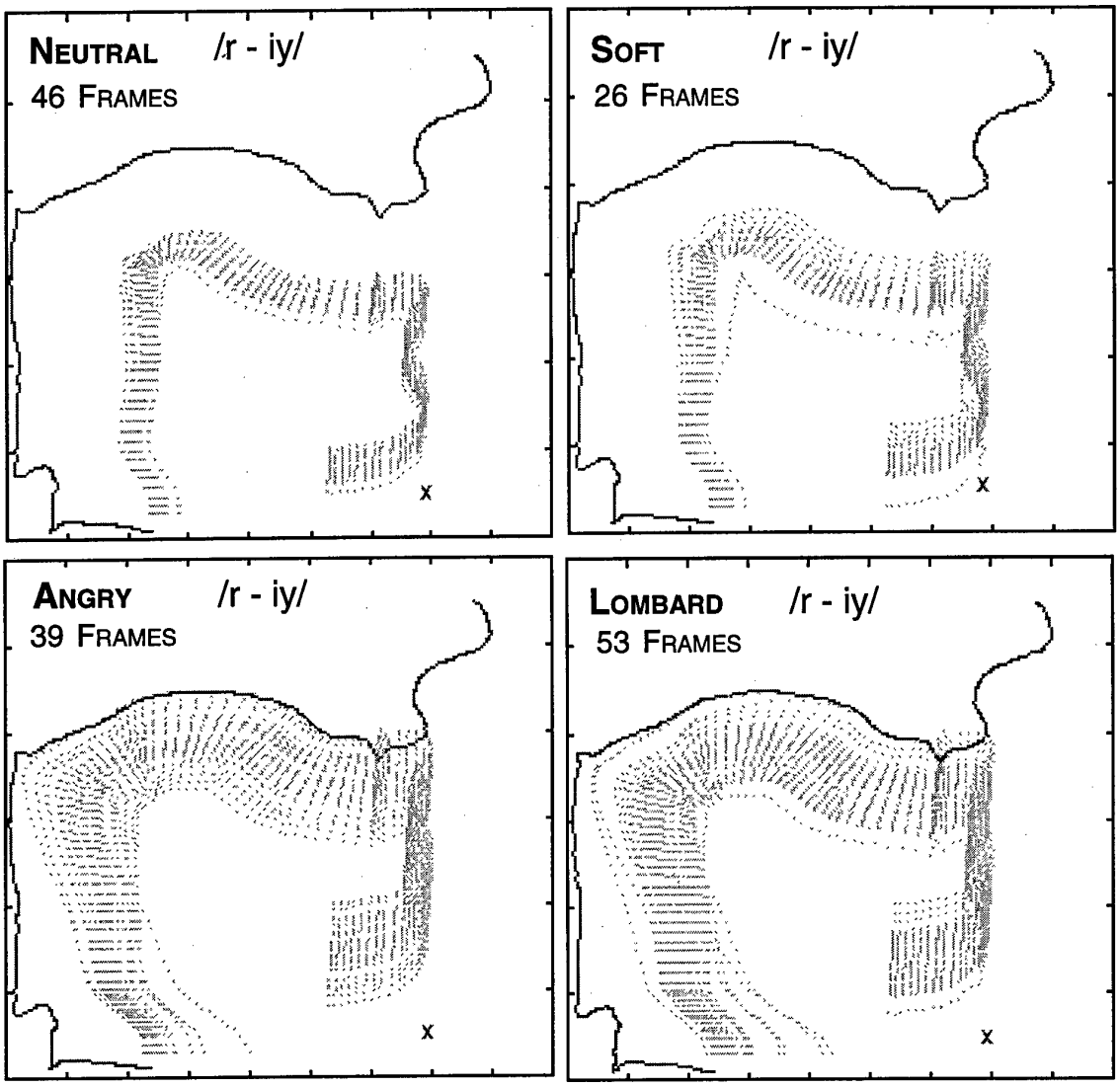


Figure 4.6: Sample vocal tract articulatory profiles for the phone sequence /r-iy/ from the utterance "freeze." Articulatory variation is shown for speech under neutral, soft angry, and Lombard effect stress conditions.

perturbation in the blade and dorsum of the tongue and the lips. This suggests that when a speaker is under stress, typical vocal tract movement is effected, suggesting a quantifiable perturbation in articulator position.

4.2.7 Analysis using 'Actual' Stressed Speech from SUSAS

One limitation of previous studies on speech under stress has been analysis of simulated stress with no confirmation using actual recordings. This problem occurs because it is difficult to obtain recordings of the same speakers in both simulated and actual stress environments. This issue was considered using speech data from the Actual Speech under Stress (roller-coaster motion-fear MF-task) and Dual Tracking Task (DT-task) portions of SUSAS (Hansen, 1998a, 1998b [61, 62]). The same five analysis areas of pitch, duration, intensity, glottal spectrum, and vocal tract spectrum were considered.

The mean of the fundamental frequency for the DT-task produced limited variations, with some speakers showing slightly increasing or decreasing mean. For DT-task, workload per-

formance was assessed using the overall RMS error for completing the acquisition tracking task [45, 46]. Speakers who experienced little variation in RMS task error also produced limited changes in the mean of the fundamental frequency. Speech from the Actual MF-task domain showed statistically significant increases in fundamental frequency mean and variance. The differences were 50–100 % larger than those observed for simulated angry or loud spoken speech, thus indicating that fundamental frequency characteristics from the simulated stress domains possess similar parameter trends (degree and direction), though actual stressed conditions were found to possess larger variations. For duration, it was determined that under Actual MF-task stress conditions, consonant and semivowel duration increased, while vowel duration decreased. Interior vowels possessed smaller decreases, and leading nasals resulted in larger duration increases than other consonants. A new quantity based on duration ratios for phoneme classes confirmed the increase in consonant duration at the expense of decreased vowel duration under MF-task stress. For the MF-task, both word and phoneme classes showed significant increases in RMS intensity with respect to neutral (+137 % to +190 %). Consonant intensity, especially nasals, also showed marked increases for mean and variance of RMS intensity. Intensity ratios suggest that there was a slight intensity shift from consonants to vowels. For the glottal source spectrum, average spectral content increased dramatically for all speakers in the Actual MF-task domain. Average spectral tilt significantly decreased for all speakers during the roller-coaster task, thus implying a large increase in the high frequency energy of the glottal pulse. Finally, for the vocal track spectrum, first formant locations generally increased for MF-task, while for the DT-task they generally decreased. Higher formant locations either remained constant, or increased slightly in frequency. First formant bandwidths normally increased, and fourth formant bandwidths always decreased. Variation in second and third formant bandwidths were mixed.

In general, variation in speech features under actual stress conditions support earlier observations from simulated stress conditions. For the Actual Stress task domain, parameter variations in general followed those observed for loud, angry, and/or Lombard styles. In most cases, variations were more pronounced (i.e., larger shifts in mean pitch, first formant location for vowels, etc.).

4.2.8 Summary from SUSAS Analysis

This section presented a brief discussion of the analysis of speech under stress. The focus was on speech from simulated stressed conditions. A similar evaluation was also conducted on speech from actual stress conditions to confirm direction and degrees of speech parameter variation. In all, well over 10 000 statistical comparisons were conducted. The interested reader may consider the following references for a more complete discussion of these results (Hansen, 1988, 1989, 1995, 1996 [54, 55, 71, 59, 60]). However, these results serve to motivate the type of speech processing needed to address recognition of speech under stress.

4.3 Analysis of Speech Under Stress using the SUSC-0 Database

A study on the analysis of speech from pilots under stress was conducted at the U.S. Air Force Research Laboratory in Rome, NY [52]. The study concentrated on the stress of an F-16 pilot while conversing with his wing man and the tower controller during an in-flight emergency. The data was taken from the NATO SUSC-0 database and consisted of approximately 10 minutes of voice communications. The study was performed to determine if a change in the Amplitude Modulation (AM) and Frequency Modulation (FM) can be detected in the pilot's voice in the vicinity of the fundamental frequency and the first and second format frequencies when the pilot was under stress. To compare the modulation characteristics of the utterances, two ratios were used. The first ratio, Es, is defined as the ratio of the energy of the AM envelope to the energy

of the bandpass filtered signal. The second ratio, fs, is the ratio of the geometric mean of the AM envelope to the arithmetic mean of the AM envelope.

The results of the study showed that both amplitude and frequency modulation around the fundamental frequency increased as stress on the pilot increased. Within the vicinity of F1 (maximum amplitude) the amplitude modulation and the frequency modulation decreased with increasing stress. Further, the Es and fs ratios decreased with increasing stress. It was also observed that Es and fs did not vary with stress in the frequency area of F2.

4.4 Selected References of Interest:

Here, we summarize several references which have considered analysis of speech under stress. The reference section at the end of this report contains all references cited in this chapter.

1. P. Benson, "Analysis of the Acoustic Correlates of Stress from an Operational Aviation Emergency," Proc. ESCA-NATO Tutorial and Research Workshop on Speech Under Stress, Lisbon, Portugal, pp. 61–64, 1995.

2. Z.S. Bond, T.J. Moore, "A note on Loud and Lombard Speech," *ICSLP-90: Proc. Inter. Conf. Spoken Lang. Proc.,* pp. 969–972, Kobe, Japan, Nov. 1990.

3. K.E. Cummings, M.A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," *J. Acoust. Soc. Am.,* **98**(1) 88–98, 1995.

4. K. Gopalan, "Voice Analysis of SUSC-0 Database," U.S. Air Force Research Laboratory Technical Report, 1997.

5. J.H.L. Hansen, M.A. Clements, "Stress Compensation and Noise Reduction Algorithms for Robust Speech Recognition," *ICASSP-89: Inter. Conf. on Acoustics Speech and Signal. Proc.,* pp. 266–269, Glasgow, Scotland, May 1989.

6. J.H.L. Hansen, O.N. Bria, "Lombard Effect Compensation for Robust Automatic Speech Recognition in Noise," *ICSLP-90: Proc. Inter. Conf. Spoken Lang. Proc.,* pp. 1125–1128, Kobe, Japan, Nov. 1990.

7. J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition," *Speech Communications, Special Issue on Speech Under Stress,* vol. 20, pp. 151–173, Nov. 1996.

8. J.W. Hicks, H. Hollien, "The Reflection of Stress in Voice-1: Understanding the Basic Correlates," *1981 Carnahan Conf. on Crime Countermeasures,* 189–195, 1981.

9. J.C. Junqua, "The Influence of Acoustics on Speech Production: A Noise-induced Stress Phenomenon Known as Lombard Reflex," *Speech Communication,* vol. 20, Nos. 1–2, pp. 13–22, 1996.

10. I. Kuroda, O. Fujiwara, N. Okamura, N. Utsuki, "Method for Determining Pilot Stress Through Analysis of Voice Communication," *Aviation, Space, & Env. Med.,* **5**:528–533, 1976.

11. P. Lieberman, S. Michaels, "Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech," *J. Acoust. Soc. Am.,* **34**(7):922–7, 1962.

12. D.B. Pisoni, R.H. Bernacki, H.C. Nusbaum, and M. Yuchtman, "Acoustic-phonetic correlates of Speech produced in noise," *ICASSP-85: Proc. Inter. Conf. Acoust., Speech, Sig. Proc.,* pp. 1581–1584, 1985.

13. R. Ruiz, E. Absil, B. Harmegnies, C. Legros, D. Poch, "Time and spectrum-related variabilities in stressed speech under laboratory and real conditions," *Speech Communication,* **20**:111–130, 1996.

14. B.J. Stanton, L.H. Jamieson, G.D. Allen, "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions," *IEEE 1988 ICASSP,* pp. 331–334, 1988.

15. L.A. Streeter, N.H. Macdonald, W. Apple, R.M. Krauss, K.M. Galotti, "Acoustic and Perceptual Indicators of Emotional Stress," *J. Acoust. So. Am.* **73**(4) 1354–1360, 1983.

16. C.E. Williams and K.N. Stevens, "Emotions and Speech: Some Acoustical Correlates", *Journal of Acoustical Society of America,* Vol. 52, No. 4, pp. 1238–1250, 1972.

Chapter 5

Stress Classification and Detection

5.1 Introduction

In this chapter we discuss commercial stress classification and assessment systems, and summarize some recent research studies on voice stress classification. While a number of voice stress classification systems are available on the market today, the scientific basis on which they are built is not well understood. A systematic objective evaluation would be a good first step to determine their usefulness. The assessment of stress in a speaker's voice is an important issue in monitoring speaker state, especially in military environments where high physical task stress or fatigue induced stress can occur, or in forensic applications for law enforcement or security applications.

Today, commercial based speech recognition systems can achieve more than 95 % recognition rates for large vocabularies in restricted paradigms. However, their performance degrades greatly in stressful situations, such as a pilot in an emergency situation, a military operator under a heavy workload, or a medical team that is exhausted due to lack of rest. Similar losses in performance occur for other recognizers such as automatic speaker recognition systems. It is suggested that algorithms which are capable of detecting and classifying stress could be beneficial in improving automatic recognition system performance under stressful conditions. Furthermore, there are other applications for stress detection and classification. For example, a stress detector could be used to detect the physical and/or mental state of a pilot and that detection could put special procedures in place such as the rerouting of communications, the redirection of action, or the initiation of an emergency plan. To be able to detect and classify stress, it is necessary to understand the effects of stress on acoustical features. Thus far, differences in acoustical features between neutral and stressed speech brought on by a variety of emotions and the Lombard effect[1] have been the focus of a number of research investigations [1, 13, 54, 60, 85, 160, 146]. We have seen in Chapter 4 that many speech production features change when a person is speaking under stressful conditions.

This chapter is organized as follows. In Sec. 5.2, traditional methods for stress classification are discussed. Most commercial based systems fall into this area. In Sec. 5.3, methods proposed in the past few years using neural network concepts are presented. These methods employ speech features derived from a linear speech model, and typically features which are cepstral-based [73, 162]. Next, Sec. 5.4 considers more recent classification experiments which use linear based speech features and optimum Bayesian detection theory. These experiments were conducted on linear features such as duration, intensity, pitch, glottal source, and vocal tract spectrum for stress classification [167]. Next, it was shown in a previous study [26, 25] that the TEO-based (Teager Energy Operator) nonlinear speech feature has the potential to improve stress classification performance. In a recent USAF study, (verified by Hansen at RSPL, Univ.

[1]The Lombard effect occurs when a speaker, in either a conscious or sub-conscious manner, modifies his speech production in order to increase his communication ability in a noisy environment.

Colorado/Duke Univ.) several TEO-based nonlinear features were found to be very effective for both stress classification and stress assessment (Zhou, Hansen, Kaiser [165, 166, 168]). Therefore, in Sec. 5.5 several new nonlinear based features are summarized which have shown promise in both classification and assessment of speaker stress. When possible, results using speech data discussed in Chapter 3 (i.e., SUSAS and SUSC-0) are presented.

5.2 Traditional Methods for Commercial Voice Stress Analysis

Traditional methods for detecting the stress in a speaker's voice have evolved from the early interest by military and law enforcement agencies in the detection of deception. Military interrogators and law enforcement interviewers were anxious for a capability that would aid in determining whether a subject was making a truthful statement or lying. Such a voice stress analysis (VSA) capability would be extremely valuable in gathering information that could impact on the outcome of a battle or a trial. The reason for this interest lies not only in its information value but in the non-intrusive and efficient way that VSA equipment promises to obtain information that otherwise could not be obtained so quickly and conveniently. While military interest has continued to increase, law enforcement agency interest has grown immensely which has created a commercial market for VSA equipment.

An additional growing civilian application of VSA technology and equipment that is creating a market is pre-employment interviews. As a result, several commercial VSA systems are available in both hardware and software. The systems range in price from approximately \$100 to \$10 000. Many of the vendors offer training courses and some of these courses are intensive and require as much as a week or more to complete.

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 [97]) 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. The determination of the degree of stress contained within a selected voice sample is made through the visual examination of the chart by a trained examiner [29]. The examiner looks for characteristic shapes related to amplitude, cyclic change, leading edge slope, and square waveform shapes called blocking.

5.3 Neural Networks with Linear Speech Model-based Features

5.3.1 Cepstral-based Features

In a previous study conducted at RSPL [73, 162, 161], a neural network based classification algorithm was considered for stress classification using cepstral-based features which have traditionally been employed for recognition. Five cepstral feature sets were investigated, which included Mel C_i (C-Mel), delta Mel DC_i (DC-Mel), delta-delta Mel $D2C_i$ (D2C-Mel), autocorrelation Mel AC_i (AC-Mel), and cross-correlation Mel $XC_{i,j}$ (XC-Mel) cepstral parameters. The first three cepstral features (C_i, DC_i, and $D2C_i$) had been shown to improve speech recognition performance in the presence of noise and Lombard effect [74]. The AC_i and $XC_{i,j}$ features were new in that they provide a measure of the correlation between Mel-cepstral coefficients.

The Mel-cepstral (C-Mel) parameters are well known as features that represent the spectral variations of the acoustic speech signal. It is suggested that such parameters are useful for stress

classification since vocal tract and spectral structure vary due to stress. The C-Mel parameters are able to reflect these energy shifts.

The DC-Mel and D2C-Mel parameters provide a measure of the "velocity" and "acceleration" of movement of the C-Mel parameters. These features are calculated by performing polynomial fitting of the C-Mel parameters and taking the derivative of the polynomial itself. This may differ from other studies which use a first and second order difference method to estimate DC_i and $D2C_i$ respectively. It appears that the reason delta parameters are more robust to stress variations is due to their reduced variance across stress conditions. This trait suggests that while these features are more useful for recognition, they may be less applicable to stress classification.

It is suggested that the two more recently derived feature representations (AC-Mel and XC-Mel) could be more successful in representing variations due to stress. The AC-Mel features are calculated as follows,

$$AC_i^{(\ell)}(k) = \sum_{m=k}^{m=k+L} \Big[C_i(m) * C_i(m+\ell)\Big] \Big/ \sup_k AC_i^{(\ell)}(k), \tag{5.1}$$

where k is the frame number, L is the correlation window length, ℓ the number of correlation lags, and i the Mel coefficient index. When $\ell = 0$, AC_i models the relative power between frequency bands. For $\ell > 0$, AC_i models spectral slope and changes in the frame to frame correlation variation due to stress. The XC-Mel coefficients are similar to the AC-Mel coefficients except that the cross-correlation is found from one Mel coefficient C_i to another C_j across frames,

$$XC_{i,j}^{(\ell)}(k) = \sum_{m=k}^{m=k+L} \Big[C_i(m) * C_j(m+\ell)\Big] \Big/ \sup_k XC_{i,j}^{(\ell)}(k). \tag{5.2}$$

The XC-Mel parameters $XC_{i,j}$ provide a quantitative measure of the relative change of broad versus fine spectral structure in energy bands. Since the correlation window length ($L = 7$) and correlation lags ($\ell = 2$) are fixed in this study, the correlation terms are a measure of how correlated adjacent frames are over a 72 ms window (24 ms/frame and 8 ms skip rate). It is apparent that both AC-Mel and XC-Mel parameters provide a measure of correlation and relative change in spectral band energies over an extended window frame. Feature analysis suggests that the AC-Mel parameters have similar properties to the XC-Mel parameters. In addition, the AC-Mel parameters can be directly compared with other selected feature sets since they are based on a single coefficient index i. Therefore, AC-Mel parameters appear to be a better choice for stress classification than the XC-Mel parameters.

5.3.2 Neural Network Classifier

A neural network stress classifier was formulated using mono-partition features (i.e., a single phone class partition). Each partition of speech features was propagated through two hidden layers of the neural network to an output layer that estimates the stress probability scores. The neural network training method employed was the cascade correlation back-propagation network using the extended delta-bar-delta learning rule [106]. This method was selected due to its flexibility, and because it is capable of forming the complex contoured hypersurface decision boundaries needed for the stress classification problem. Fig. 5.1 shows the structure of this classification system.

5.3.3 Neural Network Stress Classification Evaluations with Cepstral Features

The neural network stress classification algorithm was evaluated using a collection of features from frame- and word-level features. Both fine and broad stress classes were evaluated. The fine (i.e., ungrouped) stress classes were the 11 stress conditions from the simulated portion of

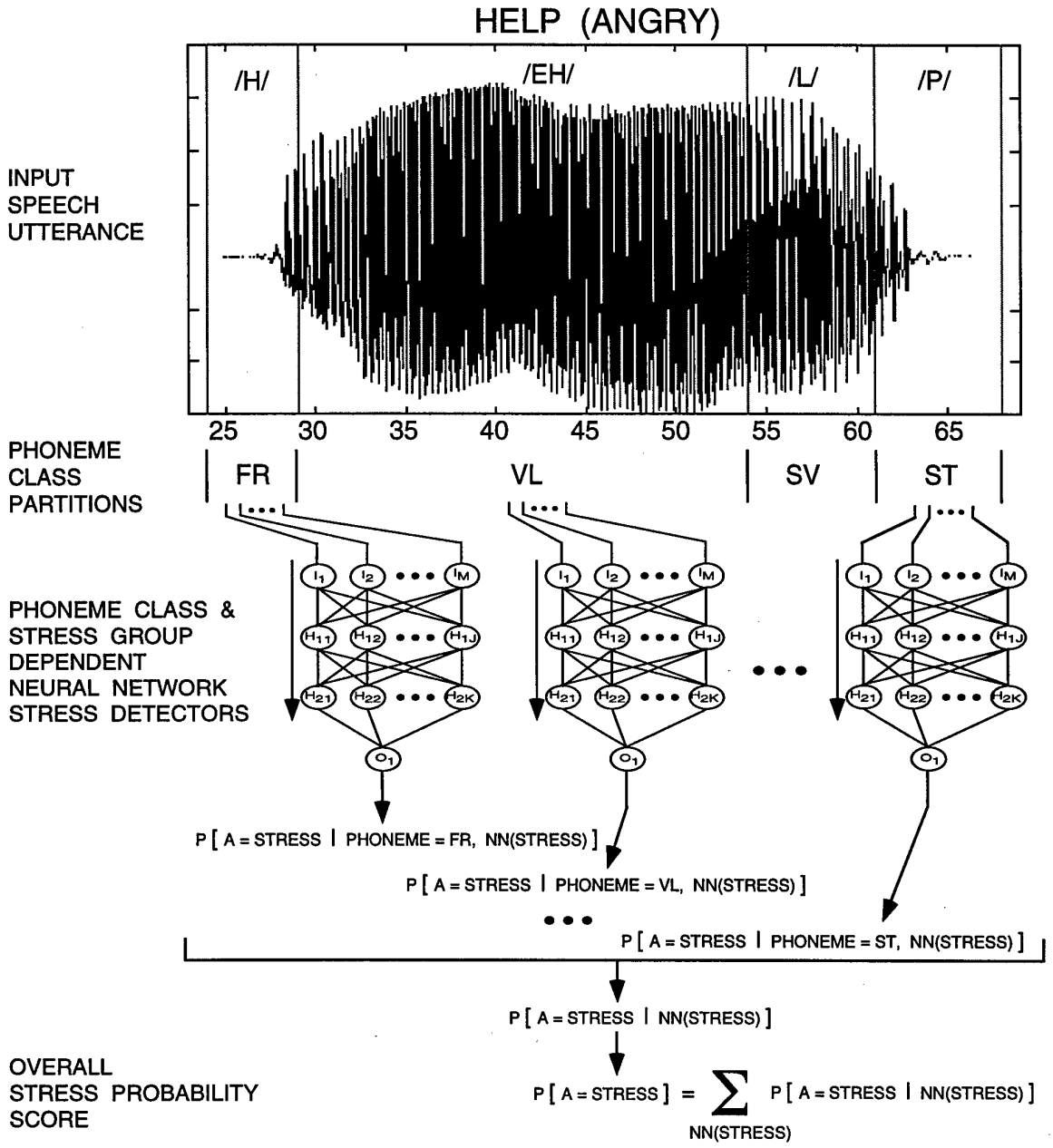


Figure 5.1: Stress classification method using phoneme based neural networks, with output scores combined for an overall stress score.

SUSAS. Ungrouped stress class neural network classifier performance is summarized in Table 5.1 using closed 35-word test sets. Classification rates ranged from **11–17 %** for the 35 word test set, which is greater than chance (i.e., **9 %**). It is clear that for some stress conditions reasonable classification performance is attained. A similar test using a 5 word vocabulary produced higher average classification rates (i.e., 32–67,%). It is also useful to point out that these results were for a 1 of N stress class test, versus a pairwise neutral versus particular stress style. Pairwise classification results are always significantly higher, since some stress conditions for SUSAS are similar (e.g., angry and loud, clear and Lombard, etc.).

Stress Classification Performance Single Speaker, 35 Words, Stress Ungrouped Closed Vocabulary Test Set				
Stress Class	Classification Rate (%)			
	C_i	DC_i	$D2C_i$	AC_i
Angry	6.20	0.00	19.49	4.96
Clear	12.50	0.00	4.27	7.32
Cond50/70	42.47	59.76	56.08	14.61
Fast	7.26	1.65	44.53	68.10
Lombard	1.64	0.00	8.53	2.54
Loud	12.40	3.73	3.15	1.69
Neutral	22.31	0.00	2.61	1.72
Question	14.05	0.00	5.15	4.76
Slow	19.01	4.76	27.56	4.31
Soft	16.53	33.09	0.00	2.38
Mean	15.44	10.30	17.14	11.24
Std. Dev.	11.33	20.12	19.62	20.35

Table 5.1: MPSC Performance for Ungrouped Closed 35 Word Test

5.3.4 Neural Network Stress Classification with Target Driven Features

Further classification studies have expanded on these neural network approaches using target driven features [162]. In this method, a wide selection of features are automatically extracted including articulatory measures, pitch, phone duration, spectral based, etc. Next, the most effective feature subset for each targeted stress condition is determined during a training phase. During classification, only those targeted features needed for a neural network stress classifier under test are employed. This allows the classifier to use the most discriminating features for classification of each stress style. A second study proposed an approach which combines stress classification and speech recognition functions into one algorithm [163]. This was accomplished by generalizing the one-dimensional hidden Markov model to an N-channel Hidden Markov Model (N-Channel HMM). Here, each stressed speech production style under consideration is allocated a dimension in the N-Channel HMM to model each perceptually induced stress condition. It is shown that this formulation better integrates perceptually induced stress effects for stress independent recognition and classification. This is due to the sub-phoneme (state level) stress classification that is implicitly performed by the algorithm. The N-channel stress HMM method was compared to a previously established 1-channel stress dependent isolated word recognition system yielding a **73.8 %** reduction in error rate.

5.4 Bayesian Stress Classification with Linear Speech Features

While neural network classifiers have shown promise, there is clearly a difference in performance based on the feature set used for stress classification. It has been shown that there are observable

differences in duration, intensity, pitch, glottal source information, and formant locations between neutral and stressed speech [54]. Therefore, it is worthwhile to evaluate their performance for stress classification, or stress detection. Here two terms, classification and detection, can be used interchangeably since only pairwise classification is considered. The methods employed for classification here are Bayesian hypothesis testing approach and distance measure.

5.4.1 Feature Description

For linear feature based stress classification, only vowel sections are extracted from the simulated domain of the SUSAS database for evaluation. The length of each vowel in msec is used as the duration feature. The intensity feature is defined as,

$$Intens = \sqrt{\frac{1}{K}\sum_{i=1}^{K} s^2(i)} \tag{5.3}$$

where $s(i)$ $(i = 1, \ldots, K)$ represents the K individual samples in the vowel. Pitch, glottal source information, and formant locations are extracted on a frame basis with frame length being 32 ms and an overlap length between adjacent frames of 16 ms. The modified simple inverse filter tracking (MSIFT) algorithm [7] is employed to extract pitch frequencies from vowel speech waveforms. Spectral slope was used as the glottal source feature. It is difficult to obtain the glottal spectral slope from the raw vowel speech waveform due to the coupling effect between the sub-glottal structure and forward portion of the vocal tract. To avoid this effect, only data obtained during closed vocal fold periods was used. This unfortunately limits the usable data. Also, it is difficult to accurately locate the boundaries between vocal fold closing and opening periods. As an approximation, a frame based log average amplitude FFT was computed versus log frequency for each vowel section.

Next, a straight line is used to approximate its envelope, and the line's slope is considered as the glottal spectral slope. Only the first two formants are used for the evaluation since the remaining formants do not show much differences between neutral and stressed speech [54]. The HTK xwaves function "formant" was employed to extract formant locations for all vowels in the SUSAS database.

5.4.2 Bayesian Hypothesis Testing versus Distance Measure Testing

A stress classifier is similar to a Bayesian hypothesis testing system. It has two hypotheses, that is, $H0$ and $H1$. Under $H0$, the speech is neutral; while under $H1$, the speech is stressed. Given an input speech feature vector, $\mathbf{x}$, $(\mathbf{x} = x_1, \ldots, x_M$; M is the vector length), the following two conditional probability densities are calculated, $p(\mathbf{x}|H0)$ and $p(\mathbf{x}|H1)$. The likelihood ratio, λ, is then defined as,

$$\lambda = \frac{p(\mathbf{x}|H1)}{p(\mathbf{x}|H0)}. \tag{5.4}$$

The decision of whether the input speech is neutral or stressed is made by comparing the likelihood or log likelihood ratio with a pre-defined threshold, β. If it is bigger than β, the input speech is labeled as stressed; otherwise it is classified as neutral. The value of β depends on what criterion is used for detection. In a stress classification system, a criterion should be selected so that the two important probabilities, the false acceptance rate (FAR) and the false rejection rate (FRR), should be as low as possible. Obviously, it is not possible to minimize both FAR and FRR, and hence, a compromise must be made between FA and FR. For some systems, the requirement for one probability is more important than the other. For a stress classification system, however, we are only interested in the overall accuracy and have no preference for either FAR or FRR. Therefore, the value of β corresponding to equal error (FAR = FRR) rate (EER)

is selected. In the experiments performed here, the values of FAR and FRR were calculated as the ratio of the number of falsely accepted vowels to the total number of vowels, and the ratio of the number falsely rejected vowels to the total number of vowels, respectively. By changing the threshold value, the value of β corresponding to EER can be found.

It is also possible to detect stressed speech from neutral by using a distance measure with prior trained feature distributions. Given an input speech feature vector, $\mathbf{x} = x_1, x_2, \ldots, x_M$; M is the vector length, two values, the distance between $\mathbf{x}$ and the neutral speech feature distribution, d_n, and the distance between $\mathbf{x}$ and the stressed speech feature distribution, d_s, are computed as follows,

$$d_n = \frac{|\hat{\mu} - \mu_n|}{\hat{\sigma}\sigma_n}, \tag{5.5}$$

$$d_s = \frac{|\hat{\mu} - \mu_s|}{\hat{\sigma}\sigma_s}, \tag{5.6}$$

where $\mu_n, \sigma_n, \mu_s, \sigma_s$ are means and standard deviations for the neutral and stressed speech features, which are obtained from training data; $\hat{\mu}$ and $\hat{\sigma}$ are the sampled estimated mean and standard deviation of the components of the input vector, $\mathbf{x}$.

This distance measure reflects how close the input test speech feature vector is to the feature distributions of neutral and stressed speech data. If d_n is smaller than d_s, the input vector $\mathbf{x}$ is labeled as neutral, otherwise, it is assigned as stressed. The distance scores can also be used to quantify the degree of stress content in the test data.

5.4.3 Linear Feature Based Evaluations

A 33 word vocabulary under neutral, angry, loud, and Lombard effect speaking styles from the simulated domain of the SUSAS database was employed for evaluations. From all identified vowels, duration, intensity, pitch, glottal spectral slope, and formant locations were extracted. For each feature, all extracted data was used to estimate the density function (*pdf*) of the feature distribution (Fig. 5.2 shows two examples, one for a Gaussian distribution for pitch and a second for Gamma distribution for glottal spectral slope for vowels under loud speaking style) to obtain ROC curves for the Bayesian hypothesis testing approach. To find average test results, the data was divided for each feature into 10 equal size sets. For each of the 10 sets, we test with one set and train with the other 9 to calculate the average EER threshold for the Bayesian hypothesis testing approach, and the mean and variance of the feature distribution for the distance measure approach.

Several testing feature vector lengths $(1, 5, 10)$ were used to obtain ROC curves and error rates. Two of the many ROC curves obtained are shown in Fig. 5.3 for stress classification between neutral and loud for mean pitch and glottal spectral slope.

Table 5.2 shows an error analysis for all five feature domains using both the Bayesian hypothesis testing approach and distance measure approach. The pairwise errors for each detection technique and feature are given for the detection of three stress conditions (angry, loud, or Lombard) from neutral speech.

Based on Table 5.2, the following observations can be made: (1) that pitch is the best feature for stress classification among the five features, (2) error rates generally decrease as feature vector length increases, (3) performance differences exist between different stress styles, and (4) mean vowel formant locations are not suitable for stress classification. The results in this section have therefore established stress classification performance using linear speech production based features with two types of optimum detection methods. Further discussion of the evaluations presented here can be found in [63].

Detection Method	Vector Length	Feature	Speaking Style of Submitted Test Speech					
			Neutral	Angry	Neutral	Loud	Neutral	Lombard
Bayesian Hypothesis Testing	1	Duration	45.13	45.38	38.21	38.72	40.77	40.26
		Intensity	40.26	37.44	34.87	32.82	40.77	39.49
		Pitch	18.95	18.57	11.94	11.63	24.08	24.18
		Glottal	33.33	36.78	41.38	41.72	42.76	42.07
		Formant 1	42.60	41.80	46.43	45.10	46.84	46.90
		Formant 2	51.48	50.88	58.20	54.51	52.98	49.88
	5	Duration	36.36	38.96	33.77	35.06	40.26	40.26
		Intensity	24.68	22.08	27.27	22.08	38.96	35.06
		Pitch	15.17	14.31	10.34	10.00	21.90	22.07
		Glottal	25.45	21.82	30.91	34.55	30.91	36.36
		Formant 1	40.60	40.30	46.12	45.82	47.91	46.87
		Formant 2	53.88	49.85	58.51	56.12	54.78	50.90
	10	Duration	41.03	35.90	38.46	35.90	38.46	46.15
		Intensity	23.08	17.95	28.21	17.95	35.90	35.90
		Pitch	12.76	11.72	7.24	8.28	20.69	19.31
		Glottal	25.00	17.86	35.71	35.71	28.57	32.14
		Formant 1	38.79	40.91	43.03	44.24	47.58	47.88
		Formant 2	55.76	47.58	59.39	57.27	53.33	55.15
Distance Measure	5	Duration	48.05	49.35	29.87	36.36	32.47	42.86
		Intensity	41.56	27.27	35.06	22.08	40.26	35.06
		Pitch	15.34	15.00	12.41	7.07	23.10	19.48
		Glottal	34.55	18.18	38.89	35.19	38.89	33.33
		Formant 1	43.58	37.76	44.63	45.97	45.82	46.72
		Formant 2	53.28	49.85	41.49	74.78	36.87	74.93
	10	Duration	43.59	53.85	28.21	41.03	30.77	46.15
		Intensity	30.77	33.33	25.64	23.08	41.03	33.33
		Pitch	14.48	12.76	12.07	4.83	21.38	17.59
		Glottal	35.71	17.86	44.44	25.93	44.44	25.93
		Formant 1	41.82	39.39	43.64	41.82	45.45	46.06
		Formant 2	54.55	49.09	40.00	74.85	38.48	76.06

Table 5.2: Detection Error Rates for Multiple Speaking Styles.

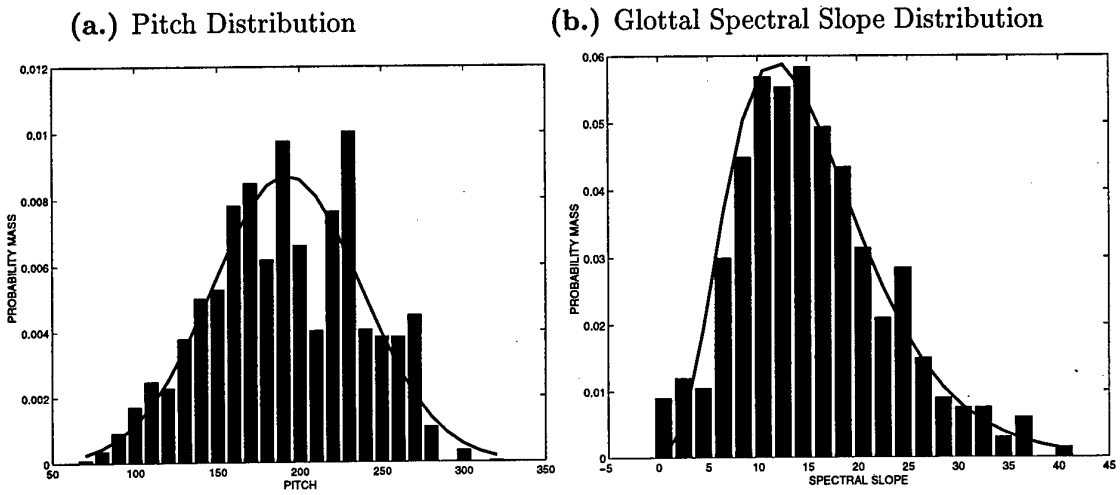


Figure 5.2: (a.) A conditional Gaussian *pdf* is used to approximate the pitch distribution of vowels under loud speaking style: $N(\mu, \sigma^2 | X \geq 0)$ with ($\mu = 192\,\text{Hz}$, $\sigma^2 = 2094$). (b.) A conditional Gamma *pdf* is used to approximate the distribution of glottal spectral slope for vowels under loud speaking style: $\Gamma(\alpha, \beta)$ with ($\alpha = 4.2329$, $\beta = 3.6612$).

5.5 Stress Classification Using Nonlinear Speech Features

In this section, recently proposed approaches to stress classification that employ on Teager Energy Operator based processing are considered. Four features are discussed, followed by evaluations using stressed speech data from SUSAS. These features have been shown to be more effective than many linear based features such as pitch and spectral structure (as reflected by MFCC parameters). Further details can be found in studies by Zhou, Hansen, and Kaiser [165, 166, 167, 168]. While some of the discussions in this section is more research oriented, the features discussed have potential as important processing tools in monitoring and assessing personnel in high stress military voice communication settings.

5.5.1 Teager Energy Operator

According to studies by Teager [152, 153, 154], the assumption that airflow propagates as a plane wave in the vocal tract may not hold, since the flow is actually separated and concomitant vortices are distributed throughout the vocal tract. Teager also suggested that hearing could be viewed as the process of detecting the energy. Based on the theory of the oscillation pattern of a simple spring–mass system, Teager developed an energy operator to measure the energy for simple sinusoids which can be believed as useful elements for speech. The simple and elegant form of the operator was introduced by Kaiser [86, 87] as,

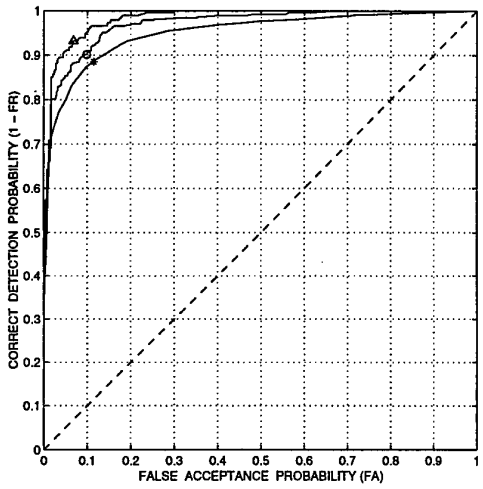
$$\begin{aligned} \Psi_c[x(t)] &= \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \\ &= [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \end{aligned} \tag{5.7}$$

where $\Psi[\cdot]$ is the Teager Energy Operator (TEO), and $x(t)$ is a single-frequency component of the continuous speech signal. Kaiser [86, 88] derived the operator for discrete-time signals from its continuous form $\Psi_c[x(t)]$, as,

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1), \tag{5.8}$$

where $x(n)$ is the sampled speech signal.

(a.) Pitch ROC



(b.) Glottal Spectral Slope ROC

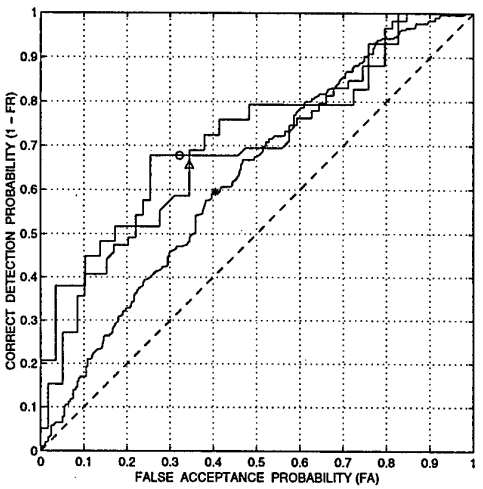


Figure 5.3: ROC detection curves for "loud" versus neutral speech using: (a.) pitch (line with $*$: input vector length is 1, EER($*$) = 11.47%; line with $\circ$: input vector length is 5, EER($\circ$) = 9.86%; line with $\triangle$: input vector length is 10, EER($\triangle$) = 6.80%). (b.) spectral slope (line with $*$: input vector length is 1, EER($*$) = 40.51%; line with $\circ$: input vector length is 5, EER($\circ$) = 32.22%; line with $\triangle$: input vector length is 10, EER($\triangle$) = 34.48%).

The TEO is typically applied to a bandpass filtered speech signal, since its intent is to reflect the energy of the nonlinear energy flow within the vocal tract for a single resonant frequency. Under this condition, the resulting TEO profile can be used to decompose a speech signal into its AM and FM components within a certain frequency band via,

$$f(n) \approx \frac{1}{2\pi T} \arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right), \tag{5.9}$$

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{\left[1 - \left(1 - \frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[x(n)]}\right)^2\right]}}, \tag{5.10}$$

where $y(n) = x(n) - x(n-1), \Psi[\cdot]$ is the TEO operator as shown in Eq. 5.8, $f(n)$ is the FM component at sample n, and $a(n)$ is the AM component at sample n [103, 104]. On the basis of this work, Maragos, Kaiser, and Quatieri [104] proposed a nonlinear model which represents the speech signal $s(t)$ as,

$$s(t) = \sum_{m=1}^{M} r_m(t), \tag{5.11}$$

where

$$r_m(t) = a_m(t) \cos\left(2\pi(f_{cm}t + \int_0^t q_m(\tau)d\tau) + \theta\right) \tag{5.12}$$

is a combined AM and FM structure representing a speech resonance at the mth formant with a center frequency $F_m = f_{cm}$. In this relation, $a_m(t)$ is the time–varying amplitude, and $q_m(\tau)$ is the frequency modulating signal at the mth formant.

Although the TEO is formulated for single-frequency signals or signals with a single resonant frequency, previous studies have shown that the TEO energy of a multi-frequency signal is not only different from that of single-frequency signal but also reflects interactions between different

frequency components [165, 166]. This characteristic extends the use of TEO to speech signals filtered with wide bandwidth band-pass filters (BPF).

5.5.2 TEO-FM-Var: FM Variation

Previous studies have shown that vowels spoken under stress generally have more instantaneous pitch variations than vowels spoken under neutral conditions. This suggests that features which represent fine excitation variations, would be useful for stress classification. To some extent, these variations are believed to be due to the effects of modulations. According to work by Maragos, Kaiser, and Quatieri [103, 104], the TEO is a nonlinear differential operator that can detect modulations in the speech signal and further decompose the signal into its AM and FM components. It is not difficult to understand that the AM-FM decomposition of a speech signal over a wide frequency band will not provide correct estimation of the real modulations. AM-FM signal analysis requires a carrier frequency which must be higher than the modulating frequencies within the signal. Because of interest in the fine excitation variations, the raw input speech is filtered using a Gabor bandpass filter (BPF) centered at the median fundamental frequency, $F0$, with a root mean square (RMS) bandwidth of $F0/2$ based on the TEO profile of the entire input. The $F0$ is estimated using the average magnitude difference function (AMDF). After the Gabor BPF, TEO analysis is performed and the resulting profile is used to separate the input speech signal into its AM and FM components using Eq. 5.9 and Eq. 5.10. A flow diagram for extracting the TEO-FM-Var feature is shown in Fig. 5.4.

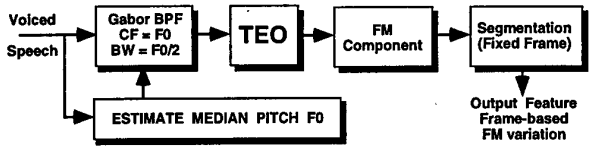


Figure 5.4: TEO-FM-Var Feature Extraction

5.5.3 TEO-Pitch: TEO based Pitch

Unlike the feature presented in 5.5.2, or the feature to be presented in 5.5.4, the TEO-Pitch feature is a direct estimate of the pitch itself. Since it is difficult for currently available techniques to correctly detect the pitch of speech under stress, especially under extreme stress, TEO processing is first applied to the raw vowel waveform. As will be explained in Sec. 5.5.4, the TEO profile has the same periodicity as pitch. Furthermore, experiments determined that it generally showed better periodicity than the raw stressed speech partly because of the square effect of TEO. Since we found that pitch usually falls within the extreme range of 50 Hz to 750 Hz (female speech from actual high stress can have pitch as high as 700 Hz), the TEO profile is bandpass filtered over (50:750 Hz) [165]. As shown in Fig. 5.5, after the BPF and segmentation, a normalized cross-correlation function (NCCF) and dynamic programming [151] is applied to detect the pitch structure. Here the waveform is first down-sampled, and candidate peaks in the NCCF are selected. Subsequently, the peaks are fine-tuned by using the NCCF of the original waveform (before down-sampling). The candidate frame-based pitch periods are determined by the average distance of two neighboring peaks within that frame. Finally, dynamic programming is employed to decide the pitch period of each frame.

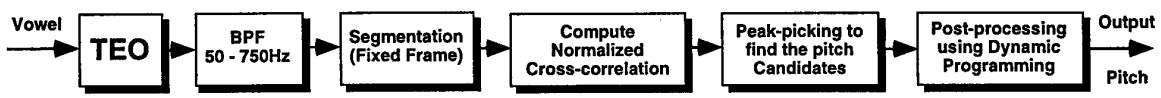


Figure 5.5: TEO-Pitch Feature Extraction

5.5.4 TEO-Auto-Env: Normalized TEO Autocorrelation Envelope Area

The third feature, named TEO-Auto-Env, also reflects the instantaneous excitation variations of speech. A flow diagram is shown in Fig. 5.6. This feature is based on the idea that the presence of stress may affect modulation patterns within the frequency bands of speech differently. It is obtained by passing the raw input speech through a filterbank consisting of 4 bandpass filters (BPF).

Each BPF output stream is processed by a TEO to estimate each profile. Our experiments show that the TEO profile of an AM-FM signal has the same periodicity as the modulating signals. Furthermore, the TEO profile periodicity is generally dominated by an amplitude modulating signal frequency. This explains why the TEO profile reflects the same periodicity as the pitch profile since both are affected by amplitude modulations. Therefore, we obtain a feature representing the fine pitch variation by analyzing the TEO autocorrelation envelope.

If we consider the fact that pitch is a slow-changing variable, we can bandpass-filter each TEO output stream through a Gabor BPF centered at the median fundamental frequency ($F0$), with the 3 dB bandwidth being roughly $F0/2$. $F0$ is obtained using the AMDF based pitch detection method on the TEO profile instead of the raw speech. Subsequently, each Gabor-filtered TEO stream is segmented into frames. In order to have equivalent averaging effects, the frame length is set to 4 times the median pitch period. Furthermore, the normalized autocorrelation function is computed for each frame. If there is no pitch variation within a frame, its normalized autocorrelation function should be a damped sinusoidal response with a straight line envelope. The area under the ideal envelope (without pitch variation) should be the same for each frame for a specified vowel, that is, $N/2$, where N is the frame length. In the case when pitch variation is present in a frame, its normalized autocorrelation envelope will not be an ideal straight line, and hence the area under the envelope will not be $N/2$. By computing the area under the normalized autocorrelation envelope and normalizing by $N/2$, it is possible to obtain 4 normalized TEO autocorrelation envelope area parameters for each time frame (i.e., one for each frequency band). This 4 parameter vector represents the TEO-Auto-Env feature per frame.

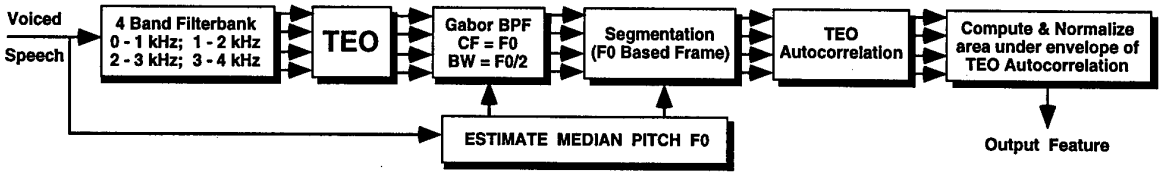


Figure 5.6: TEO-Auto-Env Feature Extraction

5.5.5 TEO-CB-Auto-Env: Critical Band Based TEO Autocorrelation Envelope

Empirically, the human auditory system is assumed to be a filtering process which partitions the entire audible frequency range into many critical bands [164]. Based on this assumption, the last proposed nonlinear feature employs a critical band based filterbank to filter the speech signal followed by TEO processing (see Fig. 5.7). Each filter in the filterbank is a Gabor bandpass filter, with the effective RMS bandwidth being the corresponding critical band.

To extract the TEO-CB-Auto-Env feature, each TEO profile of a Gabor BPF output is segmented into 200-sample (25 ms) frames with 100-sample (12.5 ms) overlap between adjacent frames. Similar to the extraction of the TEO-Auto-Env feature, M normalized TEO autocorrelation envelope area parameters are extracted for each time frame (i.e., one for each critical

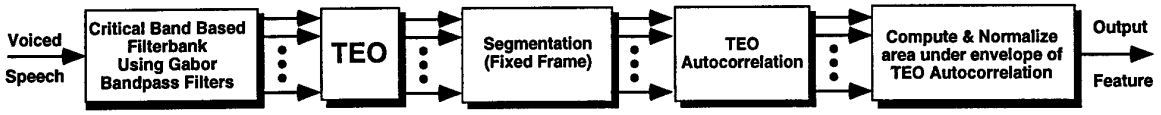


Figure 5.7: TEO-CB-Auto-Env Feature Extraction

band), where M is the total number of critical bands. This is the TEO-CB-Auto-Env feature vector per frame. Fig. 5.7 shows the entire feature extraction procedure. Since each critical band possesses a much narrower bandwidth than the 1 kHz bandwidth used for BPFs in the TEO-Auto-Env feature, post Gabor bandpass filtering centered at median $F0$ is not needed in TEO-CB-Auto-Env extraction. This makes the new feature independent of the accuracy of median $F0$ estimation.

In practice, all TEO profiles are segmented into many frames and all autocorrelation functions are normalized. As a result, the constant autocorrelation function is represented as a decaying straight line from $(0, 1)$ to $(N, 0)$, where N is the frame length. Those variations caused by harmonic distribution as well as by modulations from stress are expected to be reflected by the change in the TEO autocorrelation envelopes.

5.5.6 Evaluations

Evaluations were conducted using the SUSAS, *Speech Under Simulated and Actual Stress* database. SUSAS consists of five domains spoken under a wide range of stresses and emotions. In experiments discussed here, angry, loud and Lombard effect styles were used from SUSAS for simulated stress (speakers were requested to speak in that style; 85 dB SPL pink noise played through headphones was used to simulate the Lombard effect). Data for SUSAS actual stress was selected from the subject motion-fear domain. In the actual domain, a series of controlled speech data collection experiments were performed with speakers riding an amusement park roller coaster.

Since the TEO is more applicable for the voiced sound than for the unvoiced sound, only voiced sections of all word utterances were used for the evaluation. A baseline 5-state HMM-based stress classifier with continuous Gaussian mixture distributions was employed for the evaluations. For the purposes of comparison, the traditional pitch feature tracked by the algorithm proposed in [151] and the MFCC feature [42] were used.

The evaluation results are shown in Fig. 5.8. In general, TEO based features are effective in classifying stressed speech from neutral for both simulated and actual stress situations. Among them, the TEO-Auto-Env feature has very consistent performance across different styles of stress, but the accuracy is not as high as the TEO-CB-Auto-TEO because of fewer frequency band partitions. The TEO-CB-Auto-Env feature with fine frequency partitions, however, provide the most effective and consistent level of stress classification performance compared with MFCC and pitch information.

The evaluations in this section have shown that recently proposed nonlinear based features can be effective in the classification of speech under stress in both simulated and actual stress settings [165, 166, 167]. This assumes that the goal is to detect the presence of stress. In some military or law enforcement settings, it is also necessary to assess the level of stress in an operator's voice. The next section considers both linear and nonlinear based features for the task of stress assessment using actual emergency military voice communications between aircraft pilots from the SUSC-0 stress database.

5.6 Stress Assessment

In many military and civilian applications, it is necessary to assess whether or not a speaker is under stress. To evaluate the techniques discussed and their ability to detect real stress,

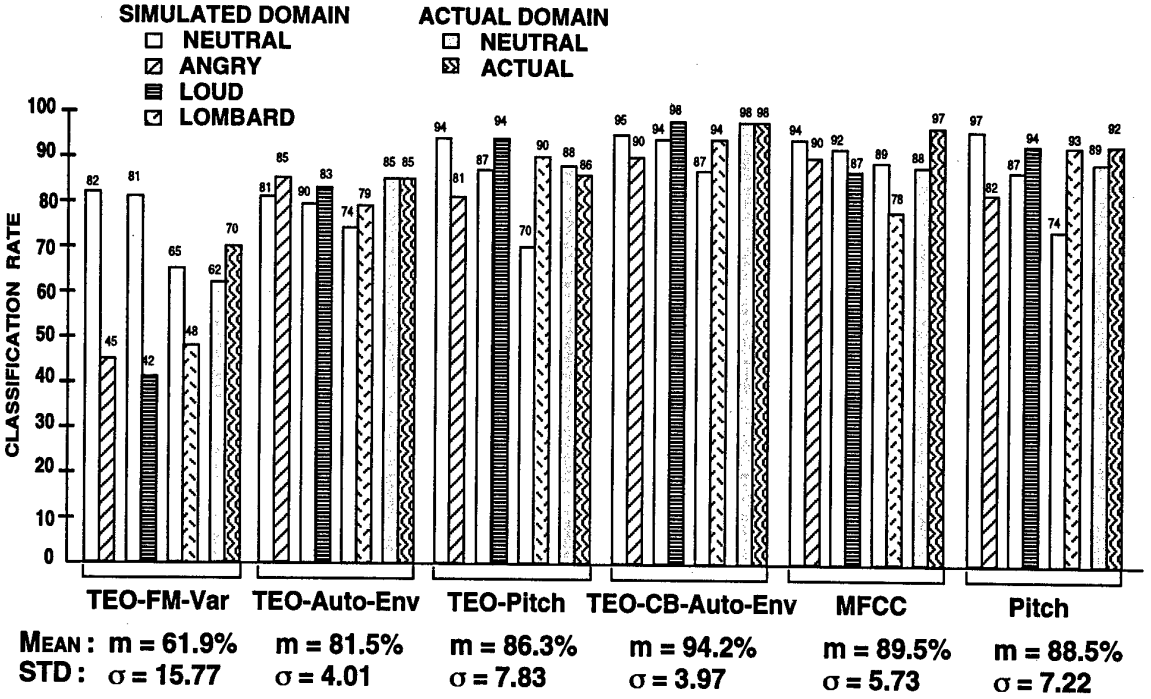


Figure 5.8: Pairwise Stress Classification Results (Mean and standard deviation of overall neutral/stress classification rates are shown; Different speaker groups were used for simulated and actual stress conditions)

the SUSC-O database containing speech of pilots under stress was processed [63]. The SUSC-0 database is from NATO IST-TG01, which consists of actual aircraft pilot communications under emergency situations. Specifically, *Mayday2* domain in SUSC-0 was used. *Mayday2* contains speech data between a pilot and controller collected from the initial ground aircraft system check, through preliminary discovery of engine emergency, until safe resolution of the emergency. The different stress degrees experienced by the pilot are reflected by his speech in *Mayday2*. A second database entitled TORONTO-AIR was also considered. This tape recording consisted of voice communications between a pilot and the controller before a fatal aircraft crashed. Since the pilot was generally unaware that an emergency was taking place until it was too late, there is only mild levels of uncertainty in his voice. Also, this tape is an 'air traffic control' (ATC) tape recorded from the ground, so high levels of noise were present. Due to these issues, the results discussed here focus on SUSC-0. Further details of the second database results are in [63]. Twelve (12) sentences from each database were extracted to represent different speaking styles for the assessment evaluation. Table 5.3 shows the 12 sentences from SUSC-0.

A baseline HMM-based stress assessor with continuous Gaussian mixture distributions was used for the evaluation. Two reference HMM models, one representing neutral speech and the other representing stressed speech, were trained. All voiced segments of word "help" under neutral conditions in SUSAS database were used to train the neutral HMM reference model. For the stressed HMM reference model, two different sets were trained, one from the simulated angry, loud, and Lombard stress conditions, and one from that actual stress roller coaster and free fall ride data, respectively. If a speech feature can assess the degree of stress regardless of text, the log likelihood ratio of the unknown speech generated by the stressed HMM model versus the neutral HMM model should be able to indicate whether it is more likely under stress or neutral. Since TEO-based autocorrelation envelope features (TEO-Auto-Env, and TEO-CB-Auto-Env), MFCC, and frame-based pitch information were shown to be very effective for stress classification, they were used to assess the stress for SUSC-0 database. Since both TEO-based

features and pitch information are only useful for voiced speech, the assessment is based on the extracted voiced portions from each utterance. To consider the variations within each utterance, 4 voiced portions per utterance (shown in Table 5.3) are extracted for the assessment.

Sentences from Mayday2 Domain of SUSC-0	
No.	Sentence
1	avionics l**IGH**t hydr**AU**lic oil pressure l**IGH**t engine indications **ARE** ...
2	**AND** you'er g**ONNA** declare an em**ER**gency or am **I**
3	... checklist **OI**l pressure malfunction **G** one-hundred ... cruise
	altitude st**OR**e jett ... throttle minimize m**O**vement ...
4	roger that **OI**l indic**A**tion is n**OW** z**ERO**
5	... **ALRIGH**t newt ... engine fault l**IGH**t still lit ... hydr**AU**lics are ...
	total p**OUN**ds six ...
6	and I'm going there and I'm there I'm des**EN**ding down to ten gr**AN**d right
	I'm n**O**t picking up a t**A**can lock
7	no I'**M** doing **ALRIGH**t now and the r**A**dial is wh**A**t
8	ok**AY** give me imm**E**diate vectors this is an em**ER**gency I'm engine **OU**t
9	g**I**ve me h**EA**dings I n**EE**d headings n**OW**
10	put the c**A**ble d**OWN** p**U**t the c**A**ble down
11	I'm h**O**t I n**EE**d the c**AbLe** ...
12	m**A**n I th**OUGH**t I w**A**s g**O**ne

Table 5.3: Sentences from SUSC-0 for Stress Assessment Evaluation. Note that bold uppercase characters represent voiced section which were used for overall stress assessment of that sentence.

The assessment results are shown in Fig. 5.9 for SUSC-0. Here, one score is obtained by finding an average output score across the four extracted voiced sections per sentence. Generally speaking, the recordings begin in a neutral relaxed setting (sentence numbers 1–2), then move into concern while pilot begins to determine the cause of the problem (sentence numbers 3–7). Finally, the pilot determines that the emergency is serious must land the aircraft without power (sentence numbers 8–11). Sentence number 12 indicates his relief after a safe landing.

Both figures ((a) and (b) in Fig. 5.9) show that the general assessment score trend is similar regardless of which anchor stress HMM reference model is used. However, the stress HMM reference model trained from actual SUSAS stress results in larger fluctuations among assessment scores. This may be because that model represents an extreme case of stress. It is noted that SUSC-0 recording can at times have high levels of background noise, so the stress assessment can be affected. We believe that it is the background noise which impacts the stress assessment because both simulated and actual stress HMM reference models produced similar results, while the actual stress HMM reference model was trained from very noisy data.

5.7 Stress Assessment and Classification Issues

We have seen that the problem of stress classification is a problem which is becoming increasingly important for military and security in the field of multi-national communications between operators/personnel. Past methods for voice stress analysis have focused on what is beleived to be microtremors in the muscles for voice production. More recent methods using digital speech processing have suggested alternative methods which offer the promise of better system integration within speech/speaker recognition or voice communications equipment for military scenarios.

While research and progress have been made in the areas of stress classification and assessment, a number of important research areas require further investigation. Here, we briefly

consider four points. First, in order to perform stress classification or assessment, two anchor models are needed (one for neutral and one for stress). These models should be trained using speech obtained from the actual stressful environments in which we wish to assess operators (i.e., aircraft pilot recordings if pilots are to be assessed). The type of stress which is displayed in one setting (aircraft cockpit), may not reflect the same workload conditions an operator may experience in another (army tank operator). Second, further research is needed to assess the consistency of stress assessment/classification for a given speaker and for unseen speakers (i.e., explore the impact of using other training data to assess new speakers). Third, there is clearly a range of emotions and workload factors which all contribute to operator 'stress.' In military scenarios an operator may experience a combination of fear, anxiety, fatigue, etc. at the same time. The ability to classify/assess this mixture of speaker traits is important in determining the stress state of the speaker/operator. Finally, there exists an unknown relationship between how computer based speech systems are able to classify stress and how humans perform stress classification. This issue is important in the collection of future databases so that better stress anchor models can be used with speech technology. From the research conducted here, it is suggested that speakers often vary how they convey stress in their speech, and that several speech features may be needed to capture the subtle differences in how speakers convey their stress state in military voice communications.

5.8 Selected References of Interest:

Here, we summarize several references which have considered classification or features related to classification of speech under stress. The reference section at the end of this report contains all references cited in this chapter.

1. P. Benson, "Analysis of the Acoustic Correlates of Stress from an Operational Aviation Emergency," Proc. ESCA-NATO Tutorial and Research Workshop on Speech Under Stress, Lisbon, Portugal, pp. 61–64, 1995.

2. D.A. Cairns, J.H.L. Hansen, "Nonlinear Analysis and Detection of Speech Under Stressed Conditions," *Journal of the Acoustical Society of America,* vol.96, no.6, pp. 3392–3400, Dec. 1994.

3. D.A. Cairns, J.H.L. Hansen, "Nonlinear Speech Analysis using the Teager Energy Operator with Application to Speech Classification under Stress," *ICSLP-94: Inter. Conf. on Spoken Lang. Proc.,* vol. II, vol. 3, pp. 1035–1038, Yokohama, Japan, Sept. 1994.

4. V.L. Cestaro, "A Comparison between Decision Accuracy Rates Obtained Using the Polygraph Instrument and the Computer Voice Stress Analyzer (CVSA) in the Absence of Jeopardy", Tech. Report, DoD Polygraph Inst., Aug. 1995.

5. J.H.L. Hansen, B.D. Womack, "Feature Analysis and Neural Network based Classification of Speech under Stress," *IEEE Trans. Speech and Audio Proc.,* vol. 4, no. 4, pp. 307–313, July 1996.

6. O. Lippold, "Physiological Tremor," *Scientific American,* vol. 224, no. 3, pp. 65–73, Mar. 1971.

7. R. Sarikaya, J.N. Gowdy, "Subband Based Classification of Speech under Stress", *IEEE 1998 ICASSP,* pp. 569–573, 1998.

8. B.J. Stanton, L.H. Jamieson, G.D. Allen, "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions," *IEEE 1988 ICASSP,* pp. 331–334, 1988.

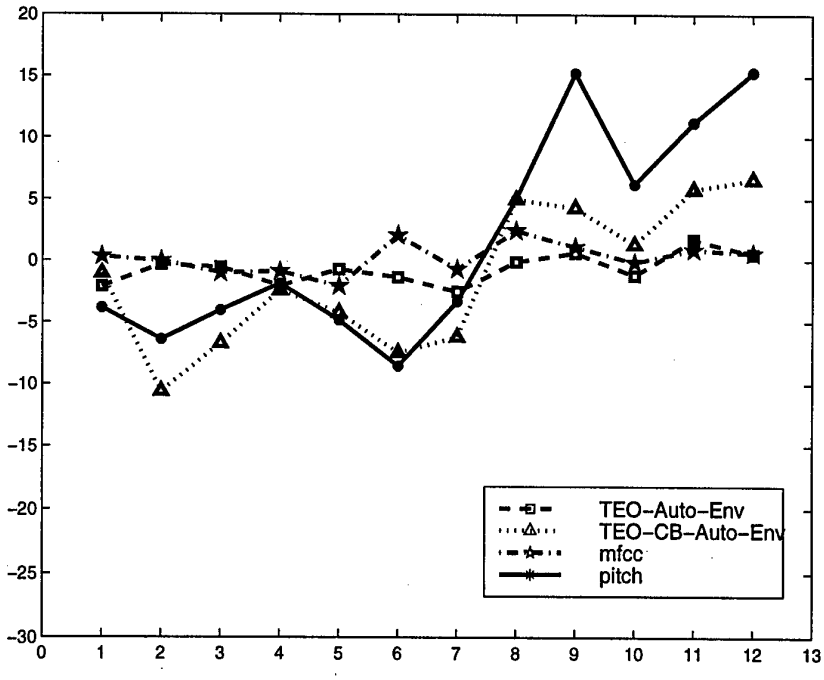
9. C.E. Williams, K.N. Stevens, "On Determining the Emotional State of Pilots During Flight: An Exploratory Study," *Aerospace Medicine,* **40** 1369–1372, 1969.

10. B.D. Womack and J.H.L. Hansen, "Classification of Speech under Stress Using Target Driven Features," *Speech Communication,* Vol. 20, Nos. 1–2, pp. 131–150, Nov. 1996.

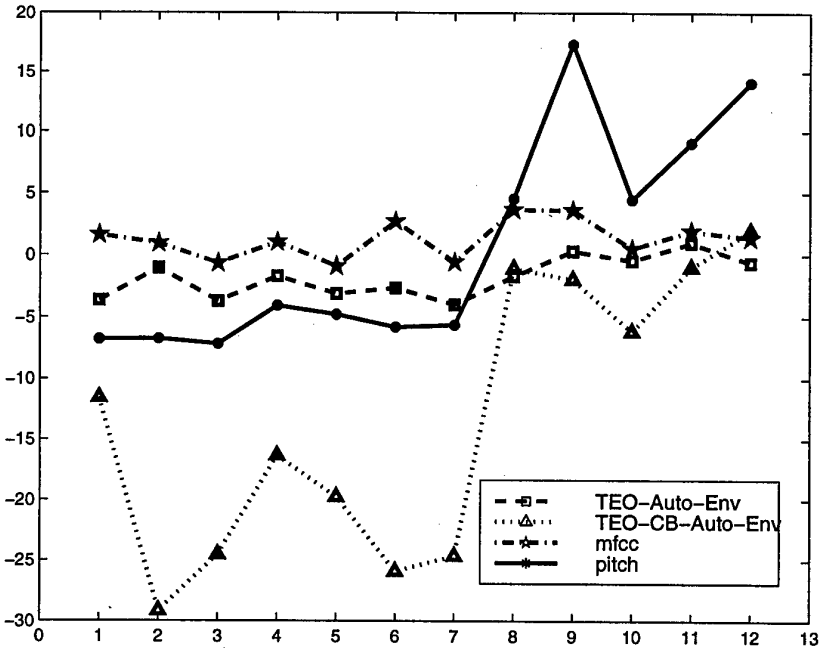
11. B.D. Womack, J.H.L. Hansen, "N-Channel Hidden Markov Models for Combined Stress Speech Classification and Recognition," accepted to *IEEE Trans. Speech & Audio Proc.*, Jan. 1999.

12. G. Zhou, J.H.L. Hansen and J.F. Kaiser, "Classification of Speech under Stress Based on Features from the Nonlinear Teager Energy Operator," *ICASSP'98,* vol. 1, pp. 549–552, Seattle, WA, 1998.

13. G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Linear and Nonlinear Speech Feature Analysis for Stress Classification," *ICSLP–98: Inter. Conf. Spoken Lang. Proc.,* vol. 3, pp. 883–886, Sydney, Australia.



(a)



(b)

Figure 5.9: Assessment results for pilot's speech from Mayday2 domain of SUSC-0 database (Log likelihood ratio is shown along Y-axis while sentence number is shown along X-axis): (a) Neutral vs Simulated stress (Loud, Angry and Lombard) HMM reference models; (b) Neutral vs Actual stress HMM reference models

Chapter 6

Speech System Evaluations

6.1 Introduction

It has been found that the robustness of various available speech technologies can be impacted by the stress level of the user of the system. Tests of speech systems have been conducted, using the stressed speech databases described in Chapter 3, at a number of research facilities around the world. The NATO RSG-10 Panel Report [149] lists the various speech technologies which are of interest to military users, however, experiments have not been carried out on all technologies using the stressed speech databases which are available. This chapter summarizes several tests of speech and speaker recognition systems as well as addresses the issues of stress and emotion for speech synthesis and speech coding systems. The related area of stress/emotion classification systems was considered in Chapter 5.

Section 6.2 describes experiments conducted on two of the databases using speech recognition systems. Section 6.2.1 summarizes experiments conducted using the DLP database using a variety of recognition systems. Section 6.2.2 looks at experiments carried out using the SUSAS database. Section 6.2.3 discusses various compensation techniques which have been proposed to improve the performance of speech recognition for stressed speech. Finally, section 6.2.4 summarizes the work done in speech recognition on the stressed speech databases.

Section 6.3 deals with tests carried out using speaker recognition systems. Since databases for speaker recognition are formulated differently than for speech recognition, only restricted evaluations were performed using the stress databases discussed in Chapter 3. Results of speaker recognition tests carried out on the SUSC-0 database are outlined in Section 6.3.1 and the results of tests carried out on the SUSAS database are given in Section 6.3.2. Conclusions on the effects of stress on speaker recognition systems are drawn in section 6.3.3.

Work being carried out on characterizing emotion in synthesis systems and coding systems is discussed in Section 6.3. At the present time, there is not as much work being carried out in this area. Here, we cite some of the recent methods proposed for imparting emotion in text-to-speech synthesis systems. Results are also presented from a recent study by Bou-Ghazale and Hansen at RSPL [19], which considers methods for imparting stress characteristics onto neutral input speech. We point out that further research is needed for a better understanding of how to effectively synthesize speech under stressful conditions, a useful task when considering military training simulators.

6.2 Speech Recognition

The issue of robustness in speech recognition can take on a broad range of problems. A speech recognizer may be robust in one environment and inappropriate for another. The main reason for this is that performance of existing recognition systems which assume a noise-free tranquil environment, degrade rapidly in the presence of noise, distortion, and stress. In Fig. 6.1, a general

speech recognition scenario is presented which considers a variety of speech signal distortions. Here, the index n represents time. For this scenario, we assume that a speaker is exposed to some adverse environment, where ambient noise is present and a stress induced task is required (or the speaker is experiencing emotional stress). The adverse environment could be a noisy military vehicle where wireless communication is used, high-stress noisy helicopter or aircraft cockpits, and others. Since the user task could be demanding, the speaker is required to divert a measured level of cognitive processing, leaving formulation of speech for recognition as a secondary task.

Speech recognition systems can be classified in a number of ways [149]. Briefly, we will consider speaker-independent (systems trained for the individual user) vs. speaker-independent (systems trained on a large population of speakers), task-dependent (systems trained on the database which is used in testing) vs. task-independent (systems trained on databases different from that used in testing) and limited vocabulary systems vs. large vocabulary systems.

The issue of robustness in speech recognition can take on a broad range of problems. A speech recognizer may be robust in one environment and inappropriate for another. The main reason for this is that performance of existing recognition systems which assume a noise-free tranquil environment, degrade rapidly in the presence of noise, distortion, and stress. In Figure 6.1, a general speech recognition scenario is presented which considers a variety of speech signal distortions. Here, the index n represents time. For this scenario, we assume that a speaker is exposed to some adverse environment, where ambient noise and a stress induced task are present (or the speaker is experiencing emotional stress). The adverse environment could be a noisy military vehicle where wireless communication is used, high-stress noisy helicopter or aircraft cockpits, and others. Since the user task could be demanding, the speaker is required to divert a measured level of cognitive processing, leaving formulation of speech for recognition as a secondary task.

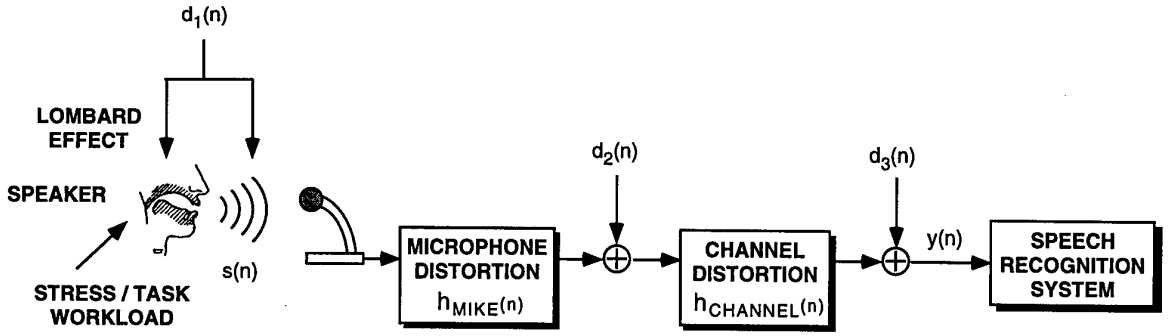


Figure 6.1: Types of distortion which can be addressed for robust speech recognition.

Workload task stress has been shown to significantly impact recognition performance [30, 54, 71, 123, 131]. Since background noise is present, the speaker will experience the *Lombard* effect (Lombard, 1911, [100, 84]); a condition where speech production is altered in an effort to communicate more effectively across a noisy environment. The level of *Lombard* effect may depend on the type and level of ambient noise $d_1(n)$ (though no studies have considered this), and has been shown to vary between male and female speakers [84]. In addition, a speaker may also experience situational stress (i.e., anger, fear, other emotional effects) or workload task stress (i.e., flying an aircraft) which will alter the manner in which speech is produced. If we assume $s(n)$ to represent a *Neutral*, noise-free speech signal, then the acoustic signal at the microphone will include distortion due to stress, workload task, *Lombard* effect, and additive noise. The acoustic background noise $d_1(n)$ will also degrade the speech signal as illustrated in Fig. 6.1. Next, if the speech recognition system is trained with one microphone and another is used for testing, then distortion due to microphone mismatch can be modeled with a frequency distortion impulse response $h_{\text{mike}}(n)$. If the speech signal is transmitted over a telephone or cellular channel, further distortion is introduced (modeled as either additive noise $d_2(n)$, or a frequency distortion with impulse response $h_{\text{channel}}(n)$). Furthermore, noise could also be present

at the receiver $d_3(n)$. Therefore, the *Neutral* noise-free distortionless speech signal $s(n)$, having been produced and transmitted under adverse conditions, is transformed into the degraded signal $y(n)$.

$$y(n) = \Big[\cdot \Big([s(n) \mid \text{workload, stress, Lombard effect}(d_1)] + d_1(n) \Big) \times \tag{6.1}$$

$$\times h_{\text{mike}}(n) + d_2(n) \Big] h_{\text{channel}}(n) + d_3(n) \tag{6.2}$$

We should emphasize that all forms of distortion identified in Equation 6.2 and Fig. 6.1 may not exist simultaneously. In studies considered in this section, the primary focus will be on speech under stress (including Lombard effect), with secondary emphasis on speech under stress with additive background noise distortion (speech data from some actual stress environments will always contain noise).

6.2.1 Speech Recognition: Tests using the DLP Database

Tests using the DLP database have been conducted using a commercial off the shelf (COTS) recognition system, a speaker-independent isolated word recognition system and speaker-independent, task-independent recognition system. The results of these experiments are summarised in the next three sections.

6.2.1.1 Commercial Off the Shelf Recognizer

Tests were conducted by the the U.K. Defence and Engineering Research Agency (DERA) using the DLP database to test a COTS speech recognition system on the DLP stressed speech database. As part of another task, a Marconi Speech and Information Systems (MSIS) speech recognition card was available for use with a speaker independent alphanumeric model set provided. The MR8 contains an implementation of the IMELDA technique to improve robustness to speaker variability, which should include variability due to stress. The recognizer was tested on the DS development set of speech files. The results show a small word accuracy fall-off from 95.1 % in the slow condition to 93.5 % in the fast condition suggesting that the IMELDA transform does provide an improvement in robustness to speaker stress.

Initial recognition experiments have been carried out on the DLP database to investigate the affect of user stress as caused by a time constrained task on speech recognizer performance. The use of a NASA-TLX performance evaluation questionnaire indicated that there was a measureable increase in user stress for the fast data display conditions. As was expected, the performance of a speaker-independent, sub-word based speech recognition system was found to be degraded during the higher stress task, as stress related speech affectations were encountered, such as higher co-articulation rates, slurring and false starts.

This experiment does confirm that cognitive stress, whether due to conflicting parallel tasks as reported in other works, or due to a time constrained task will degrade speech recognition performance. This has to be borne in mind when speech recognition applications are being considered. It will be necessary for non-adaptive recognition systems to be used in situations where user stress is likely to be irrelevant, such as non-safety critical systems. It is likely that shortcomings in recognizer performance, such as high error rate or slow working speed may cause an increase in user stress, exacerbating the problem and disenchanting the user - a serious implication for future market penetration.

Adaptive recognition techniques would be expected to cope better with increasing user stress level. It has been noted that the IMELDA technique provides good robustness against short term speaker variability and initial results with a recognizer incorporating this technique do show a less dramatic performance fall-off than with the ASTREC recognizer. Further investigations into robust recognition systems may provide further techniques to cope with stress induced degradation, as well as other performance degrading environments.

6.2.1.2 Speaker-independent, isolated word recognition system

A set of speaker-independent recognition evaluations of the DLP Database was conducted by Bou-Ghazale and Hansen of the Robust Speech Processing Lab (RSPL), Univ. of Colorado/Duke Univ. Although under high task conditions spoken errors may be present, it was assumed that the speaker had correctly uttered the individual words. When a person incorrectly utters a word or skips a word completely, the recognizer cannot recover from such human errors unless some rules or a priori knowledge are integrated within the system. For this reason, this study was intended to ignore human errors, and focus on recognition system errors caused by cognitive workload stress.

Background: The number plates were presented with two different rates: fast and slow. The resulting speech is referred to here as *high* and *moderate* task speech. The task of dictating the car plates at a fast rate was believed to be a source of cognitive stress to speakers. One point to note, the occurrance of the individual alpha numerical characters and digits is not evenly distributed for all words. For example, while in some instances, twenty-eight tokens of the word *eight* exist, there are only three tokens of the word *lima.*

Data Preparation: For the purpose of this experiment, the data was first downsampled from 20 kHz to 8 kHz. The down-sampled continuous sentences were then parsed into isolated words by using the orthographic transcription provided with the database. A word which has less than four training tokens per speaker and style was excluded from the evaluations. Four tokens of each word under each style were then randomly selected for training a speaker independent isolated word recognizer. Prior to training, the isolated tokens were played to a human listener in order to spot and correct any labeling errors such as incorrect boundaries, or incorrect label files. The following twenty-seven words were used in the evaluations: *alpha, charlie, delta, echo, eight, five, four, foxtrot, golf, hotel, juliet, kilo, nine, november, one, oscar, papa, romeo, seven, six, three, two, uniform, whiskey, x-ray, yankee,* and *zero.*

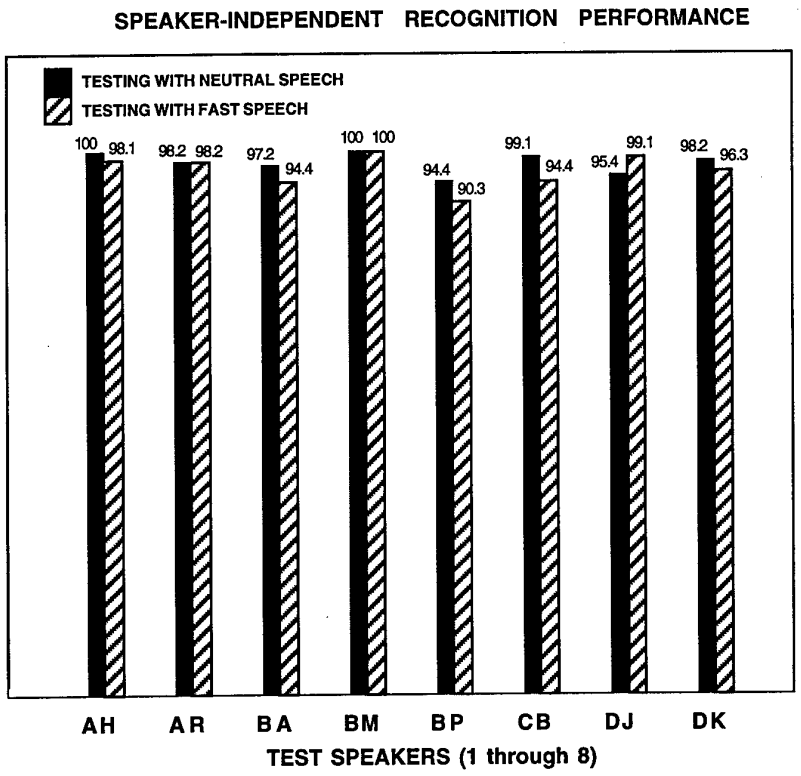
Isolated Word Training: The data was parameterized using 12 mel-frequency cepstral coefficients. A 25 ms Hamming window was used, and a first order preemphasis was applied to the data using a coefficient of 0.97. The zeroth cepstral coefficient served as the energy component. Cepstral mean normalization was performed on the cepstral parameters to compensate for long-term spectral effects. In addition, the cepstral parameters were re-scaled using a cepstral lifter as follows:

$$c'_n = (1 + \frac{L}{2} \sin \frac{\pi n}{L})\, c_n,$$

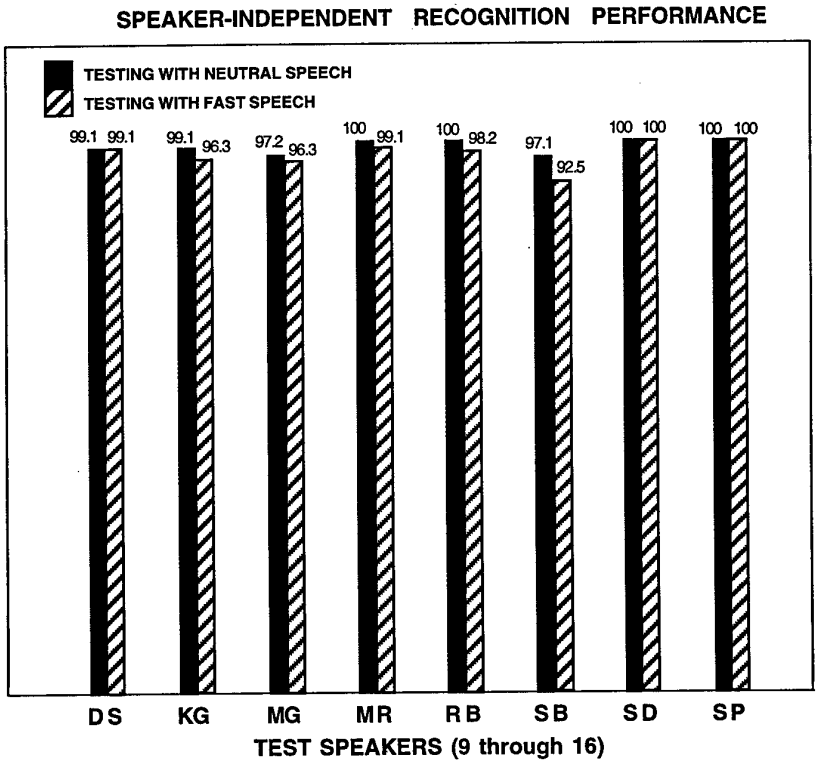
where L was set to 22. Delta and acceleration coefficients were also computed and appended to the MFCC coefficients since the performance of a speech recognition system can be greatly enhanced by adding these time derivative features.

A total of 27 words were used for training the recognizer. The speaker population consisted of 12 males and 4 female speakers. A round robin training and testing scenario was employed in order to test all speakers. Hence, a total of 60 tokens per word (4 tokens per word $\times$ 15 speakers) per style were employed for training a 2-mixture continuous density 5-state left-to-right HMM model.

Recognition Performance: A total of 1728 tokens per style were employed for testing the speaker and gender independent recognizer (results shown in Figure 6.2). The recognition accuracy of the models trained and tested with moderate task speech was 98.4 %. Models trained with moderate task speech, and tested with high task speech achieved a 97.0 % recognition accuracy. Hence, the recognition error which may be due to cognitive stress is 1.4 %. A model trained and tested with high task speech achieved a recognition rate of 97.7 %.



(a)



(b)

Figure 6.2: Individual Speech Recognition Results from DLP speech data (a) speakers 1–8, and (b) speakers 9–16.

Models Tested With	Models Trained With	
	Moderate Task	High Task
Moderate Task	98.4 %	97.5 %
High Task	97.0 %	97.7 %

Table 6.1: Speaker-independent recognition results from speech produced under moderate and high task conditions (DLP reading task). Round robin training and testing was conducted using speech spoken by a total of 16 (12 male and 4 female) speakers.

6.2.1.3 Speaker-independent, task independent recognition system

Tests were conducted by the the U.K. Defence and Engineering Research Agency (DERA) using the DLP database to test a speaker-independent, task-independent recognition system on the DLP stressed speech database.

The recognition experiments were carried out in two stages. The first set of experiments involved testing the ASTREC-216 recognizer on all the DLP database using digit and ICAO alphabet models extracted from the speaker-independent air reconnaissance mission (ARM) model set. This was to provide initial, rough estimates of the performance variations likely to be encountered in a naively implemented system with little error recovery. The second set of experiments involved a more detailed investigation of the maximum performance that could be achieved by the recognizer. As this involved much more experimentation, only speech data from a single speaker, DS, was used as a development set. The best performance measured was equivalent to that reported for more general speaker independent results.

The Recognition System: The ASTREC-216 recognizer is an implementation of a one pass, fully continuous, sub-word hidden Markov model (HMM) based recognition system. The speech signal is passed through a critical band spaced, 27 channel filterbank analyzer running in software before input to the recognizer. The simple filterbank vector is transformed into various possible representative feature vectors by a preprocessor stage of the recognizer operating at run time. The preprocessor options specified variable frame rate analysis at Euclidean distance threshold 1100, frame throw away limit of 50, then reduction to eight cosine coefficients and then time differenced.

The alphanumeric model set used was a subset of the speaker-independent ARM model set, built using a decision tree approach to provide context sensitive models and nominal task independence. The training set for male speaker models was based on the SI89 speech corpus. and comprised three read ARM reports each from 61 speakers. The female training set comprised three read ARM reports each from 61 speakers All models used were multi-state, single Gaussian HMMs, with re-estimation being carried out in the transformed domain specified by the preprocessor options. These model sets are not completely independent of the number plate reading task as in the ARM reports, alphanumerics are generally read as a continuous string rather than having phonemic contexts embedded in other words, improving the potential model performance at tasks such as this.

Initial Experiment: The initial experiments involved the running of the ASTREC recognizer on all the speech files. For this initial trial, the modifications to the model set made to accommodate the new recognition task were simply to alter the active vocabulary to exclude all but the ICAO alphabet and digit model combinations and the triphone model combinations to explain the out of vocabulary words identified from the recordings. This method of handling the out of vocabulary words was not considered to be optimal due to the task independence of the models. It would be expected that this treatment would lead to a higher insertion rate than could be

achieved with a more careful consideration of errors, however, this method would provide initial estimates of the recognizer performance, as well as giving pointers to requirements for future error handling algorithms.

Results: The results for the initial recognition experiments are summarized in Figure 6.3. The words correct performance figures include the percentage of the words actually spoken that were correctly recognized while the word accuracy figures are the percentage of displayed tokens correctly recognized. Figure 6.3(a) summarizes the words correct and Figure 6.3(b) word accuracy figures for all the speakers. The recognizer performance shows large variations across the speaker population. Previous work has reported a word accuracy figure of around 75 % for speaker-independent operation on controlled recordings on the ARM task. The performance shown here is significantly below that on the slow data rate speech for the easier task of alphanumeric recognition.

This poor result could be explained by the lack of careful level monitoring during recording sessions. Another possible source of reduced accuracy was a pollution of the recordings by power line (50 Hz) pickup from the low power fluorescent lighting used in the recording booth. This hum was more noticeable on some of the recordings than others.

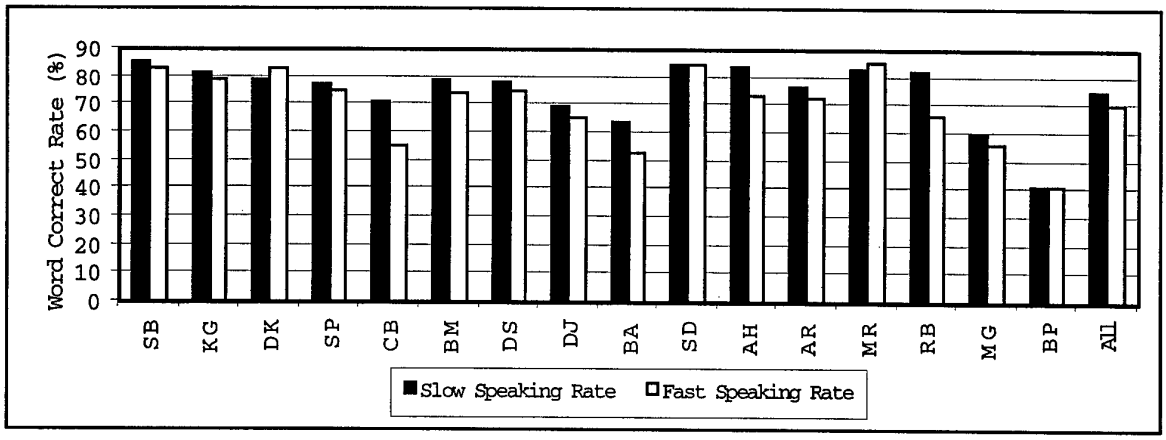
Speaker BP, on which the recognizer performed the worst, had a strong Scottish accent, which would be expected to perform less than satisfactorily against the model set which was trained on southern British English accented speech. Of the other poor results, speakers BA and MG had varying strengths of Scottish accent, while speaker DJ had a mild Welsh accent. The best performance, speaker SB, may be explained by the fact that her speech was in the data used to train the speech models.

Variation of Performance With Stress Level: Comparing the variation of performance with the change from slow to fast data rates shows a small but significant degradation as the speaker stress level, as marked by the NASA-TLX test, increased. This is broadly in line with expectations, though the figure is made less certain due to the large variations between speakers. Average variations in the words correct rate and word accuracy rate for all the speakers are 12.3 % and 20 % respectively.

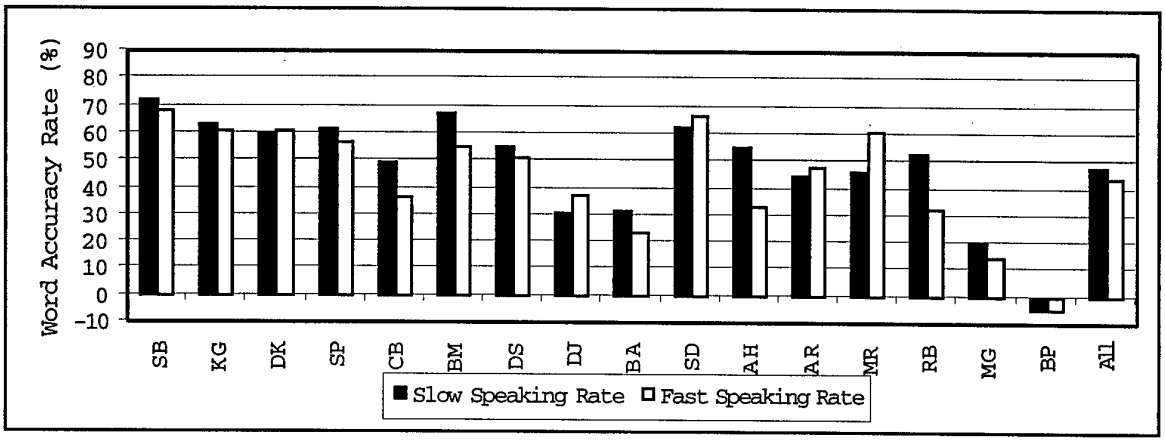
Improving The Recognizer Performance: The initial results reported above for alphanumeric recognition using the speaker independent ARM model set were fairly poor. There are several possible reasons beyond those discussed above. One was the lack of proper handling of out of vocabulary words. For the above experiments, out of vocabulary words were modeled by concatenation of sub-word units, but the performance of this was found to be lacking. Out of vocabulary words could be modeled either by a babble model or a wildcard model, or explicitly ignored from the scoring. Also, the recognizer parameters could be optimized and a syntax added. To carry out this matrix of tests on the full data set would have been impractical, so a single speaker set, DS, was chosen as a development set, due to the speaker's mild southern British English accent matching that of the model set and his high spoken accuracy.

The recognition results for speaker DS for the various experiments carried out are shown in Figure 6.4. The first experimental condition was a repeat of the initial experiment. The model set used was the male, speaker-independent ARM model set with only the models for the alphanumeric words, silence models and triphone model expansions of out of vocabulary speech.

It was noted that the models used for explaining the out of vocabulary speech were causing the majority of the insertion and mismatch errors. The second experiment simply removed the models to explain the out of vocabulary speech. It would be expected that this would allow insertion errors when a non-vocabulary word was spoken, though for speaker DS, these were few in number. The results show a significant improvement in recognizer performance as the

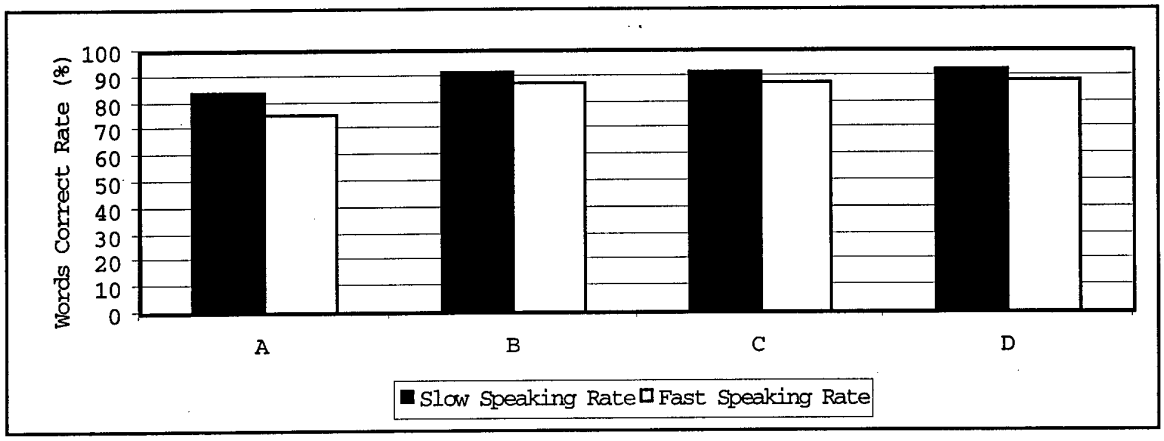


(a)

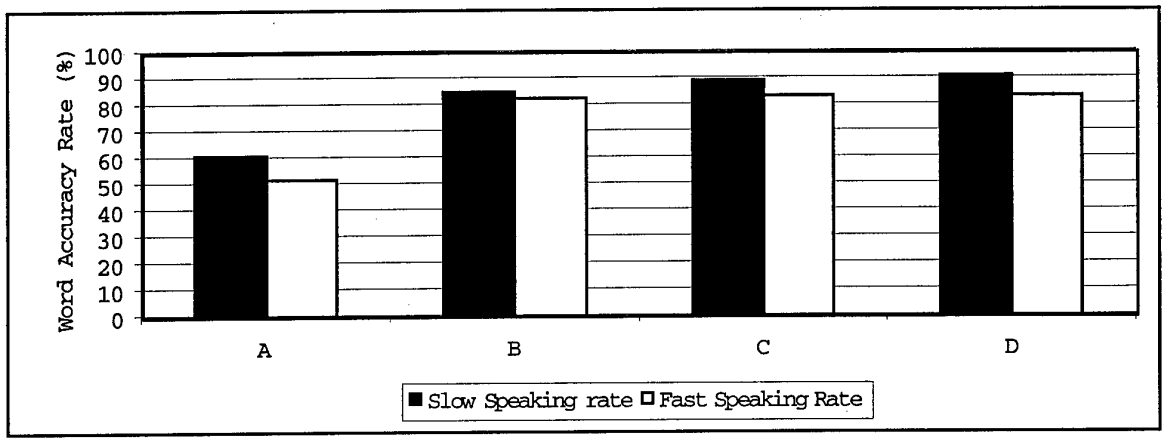


(b)

Figure 6.3: Words Correct Rate (a) and Word Accuracy Rate (b) as a Function of Visual Data Rate for all Speakers.



(a)



(b)

Figure 6.4: Words Correct Rate (a) and Word Accuracy Rate (b) as a Function of Visual Data Rate for Speaker DS. Experimental Conditions Were A: Initial Experiment, B: Use Target Vocabulary and Four Silence Models Only, C: Addition of Wildcard Model for Out of Vocabulary Speech, and D: Out of Vocabulary Words Ignored in Scoring.

poor non-vocabulary models could no longer degrade the result. The insertion rate due to out of vocabulary speech was not a significant problem.

Next, a wildcard model was included. The wildcard model returns the same Euclidean distance for each input speech observation. If the speech was in vocabulary, the correct speech model should prove a better match than the wildcard. However, out of vocabulary speech will match poorly with the speech models but the same with the wildcard, leading to the wildcard model being chosen as the best match. This technique provided a small improvement in performance indicating that the wildcard model was working as specified.

The final result in Figure 6.4 shows the performance of the recognizer on the target vocabulary only. This was achieved by specifying, in the scoring program, that certain out of vocabulary words listed in the master annotation file were to be ignored, and any word label output by the recognizer at that point was to be ignored. This provided a small performance increment over the use of a wildcard model which suggests that out of vocabulary speech was not a problem with speaker DS.

Comments on Results: The average word accuracy values from this last experiment were 91.0 % for the slow condition and 83.7 % for the fast condition. This level of performance was that

which would be expected for speaker independent recognition of a small vocabulary size. The fall-off in recognition performance with the increasing stress level was quite marked indicating that stress effects do cause a problem for speech recognition. This 7.3 % degradation in word accuracy was caused by an increase in the errors due to mismatch between input speech and the models used in the recognizer, which was expected, as slurring and co-articulation degrade the quality of the shorter digit labels.

6.2.1.4 Discussion

The results for all the experiments carried out on the DLP database are summarized in Figure 6.5. In general, there was a decrease in recognition rate as the stress level was increased, i.e., from the slow rate to the fast rate. The one exception to this was the RPSL tests with training done on the fast rate data. This result is expected since the models produced would best match the training data. As is expected, the results also show a decrease in recognition rate when going from a system trained on the database (the RPSL and MR8 results) to a task-independent system (the DERA ASTREC results).

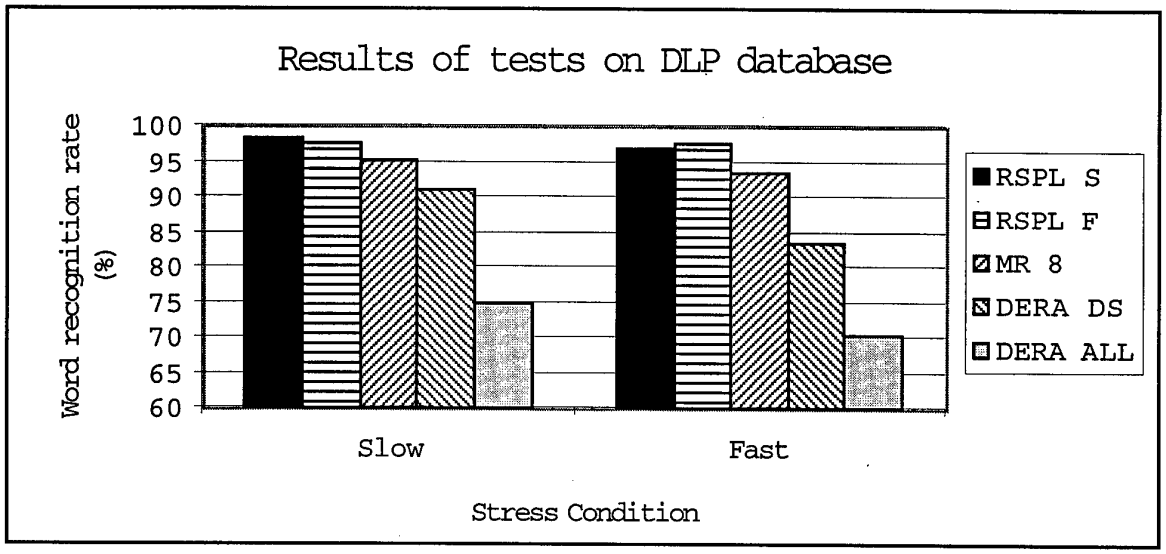


Figure 6.5: Word Recognition Rate for All tests Conducted on the DLP Database.

6.2.2 Speech Recognition: Tests using the SUSAS Database

The following subsections consider results from evaluations using the SUSAS database. The first is a study utilizing a phone recognizer. The next is two studies using isolated-word methods. The other sections summarize results of large vocabulary, continuous speech recognition systems, one of which is a COTS product.

6.2.2.1 Monophone recognition system

Large vocabulary continuous speech recognition (LVCSR) systems use low-level sub-word recognition along with higher-level language and grammar level constraints, which can at times correct sub-word and word-level recognition errors. An evaluation was conducted at RSPL, Univ. of Colorado/Duke Univ. to explore monophone recognition performance across stressed speaking conditions using the SUSAS database.

In order to investigate the effects of stress on the monophone recognition task, an experiment was performed using three stress conditions: angry, loud and Lombard as well as neutral. The

number of phones obtained from the training portion of SUSAS was not sufficient for full coverage of a standard phone set such as that used in the TIMIT speech corpus. The phone set for this experiment was limited to 29 phones. Since the SUSAS vocabulary set is limited to a 35 word vocabulary, the effects of phoneme context in the recognition task could not be considered.

The phone set is given in the Figure 6.6 along with individual recognition results. For each phone there are 12 neutral training tokens. The models were generated by round-robin training for each phone and for each speaker. As a result the neutral condition was tested with 12 tokens whereas only two tokens were available for the other three stress conditions. The ensemble of results over all nine speakers were averaged to obtain reliable recognition rates. Hence the results listed in the table comprise 108 testing tokens for neutral and 18 tokens for each stress condition.

The speech data was sampled at 8 kHz and analyzed at a skip rate of 5 ms with 15 ms window length. Ten static MFCCs were supplemented with deltas, energy and delta energy coefficients, resulting in a feature vector size of 22. Here, 3-state left-to-right HMMs with 2 mixtures were used for modeling. Several observations can made from the results. First, diphthongs, vowels and semi-vowels yielded very high recognition rates which were usually long in duration in neutral testing. Consonants rates were not as high as high energy voiced phones, with particularly low performance for stops which have high variability and short duration.

The overall recognition rate for angry speech resulted in dramatic reduction in recognition performance 30.7%, followed by loud and Lombard. The results indicate that these stress conditions resulted in tremendous variability in the acoustic parameterization which were not represented in the underlying neutral model.

A number of more detailed experiments were performed by RSPL on stressed monophone recognition to investigate the effect of phoneme positions (initial, middle, final) on the stressed speech recognition rates. Observations on the effects of phoneme locations together with the some of the results for stressed monophone recognition are as follows:

1. For nasals, recognition rates increase in general for all stress conditions if the phoneme comes in the final position instead of initial position. The differences between scores are dramatic (e.g., /m/ in "mark -vs- histogram", and /n/ in "no -vs- change -vs- on").

2. For voiced stops (/d/, /g/), the rates increased across all stress conditions when they come in the middle or final positions rather than the initial position (e.g., /d/ in "degree-wide" and /g/ in "go-degree").

3. For dipthongs (/ey/, /aw/, /iy/), while scores for neutral were not sensitive to phoneme position, the recognition scores all increase for the other stress conditions, when they come in the middle or final position versus when they come in the initial position.

4. For /th/, recognition rates increase across all stress conditions if the phone comes in the initial position versus the final position.

Change in Recognition Rate for /th/ in Initial vs. Final word position.

	Neutral	Angry	Loud	Lombard
Difference:	+3.7%	+27.8%	+22.3%	+16.6%

5. For /f/ phone, recognition rates increase in general across all stress conditions if /f/ comes either at the middle or final position.

6. For /z/, recognition rates increase for neutral, angry and Lombard if /z/ comes in the final position rather than initial position. It decrease for loud speech if it comes in the final position.

Change in Recognition Rate for /th/ in Final vs. Initial word position.

	Neutral	Angry	Loud	Lombard
Difference:	+18.5 %	+44.5 %	−16.6 %	+16.6 %

7. For /p/, while there is very little change in recognition rates for the neutral case for different phone positions, recognition rates increased dramatically for the other stress conditions when /p/ comes in the final position rather than initial position.

Change in Recognition Rate for /p/ in Final vs. Initial word position.

	Neutral	Angry	Loud	Lombard
Difference:	−6.5 %	+33.4 %	+44.4 %	+16.7 %

8. For /k/, recognition rates increase across all stress conditions when the phone comes in the middle rather than in the final position.

Change in Recognition Rate for /k/ in Middle vs. Final word position.

	Neutral	Angry	Loud	Lombard
Difference:	+26.9 %	+16.7 %	+11.1 %	+16.6 %

6.2.2.2 Speaker-dependent isolated word systems

This section considers results from evaluations using the SUSAS database. Two studies focus on isolated-word methods, one conducted by Grupo De Technologa Del Habla (GTH) in Spain and the other by the Robust Speech Processing Lab (RSPL) in the USA.

6.2.2.3 Test conducted by GTH

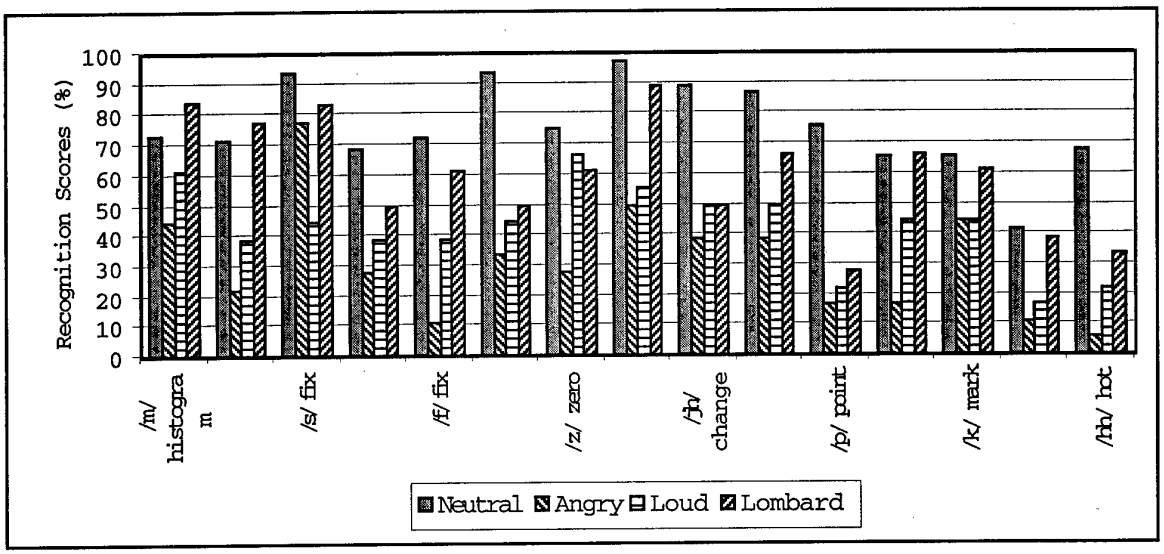
The Automatic Speech Recognition (ASR) System: The recognition tests were performed with an isolated word recognizer developed by Grupo De Technologa Del Habla (GTH), Universidad Politécnica De Madrid, a standard discrete HMM system using one model per digit in a speaker-dependent task.

The front-end of the ASR system used the following features: 10 mel-cepstrum coefficients and the average energy extracted from each frame. The MFCC were obtained using a bank of 17 mel-scaled, triangular band-pass filters applied to the DFT spectrum computed from a frame of 256 points windowed by a Hamming window with preemphasis. The sampling rate was 8 kHz and the frame advance rate was 80 frames per second. Dynamic parameters were not extracted, so, only 11 coefficients represent each frame. The process of quantization used only one codebook composed of 128 centroids.

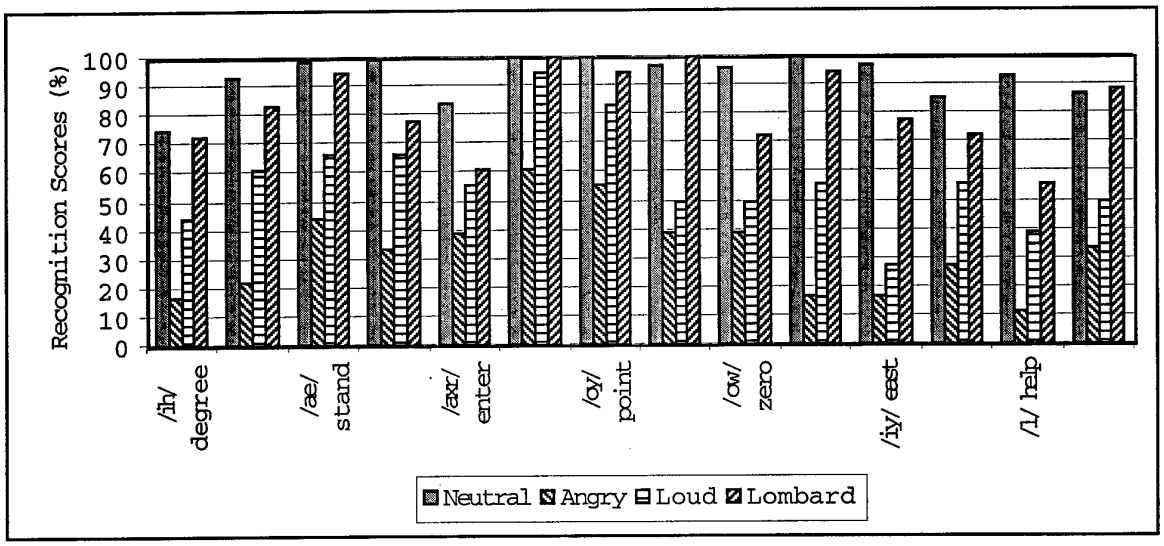
Every speech file, in both the training and test sets, was automatically segmented in order to detect the beginning and the end of the utterance. For each digit, six frames of silence were added at the beginning and the end of the isolated word for modeling the word and the silence simultaneously. For each word, a discrete HMM with six emitting states was trained using the "train" data only. The two models of silence were trained in the same way using a discrete HMM with three emitting states. The recognition stage used a standard Viterbi decoder for isolated speech with initial and final silence models appended to each word.

Experiments were carried out on the talking styles and single tracking task domains from the SUSAS database. A complete report of these experiments is available [50].

Results: Figure 6.7 summarizes the average recognition rate for all speakers under the conditions in the talking styles domain (neutral, angry, clear, fast, loud, question, slow and soft). These conditions are simulated stress conditions. Note: results of recognition tests using the



(a)



(b)

Figure 6.6: Monophone Recognition Rates for Various Stress Conditions for (a) Consonants and (b) Vowels, Diphthongs and Semi-Vowels.

training portion of the database are included in order to evaluate the correct performance of the recognizer.

Figure 6.8 summarizes the average recognition rate for all speakers in simulated task stress conditions: moderate stress (cond50), high stress (cond70) and Lombard conditions (lombard). Also shown are the results for neutral speech in order to compare the performance of the ASR system with the other stress conditions and the results for the training set in order to evaluate the correct performance of the recognizer.

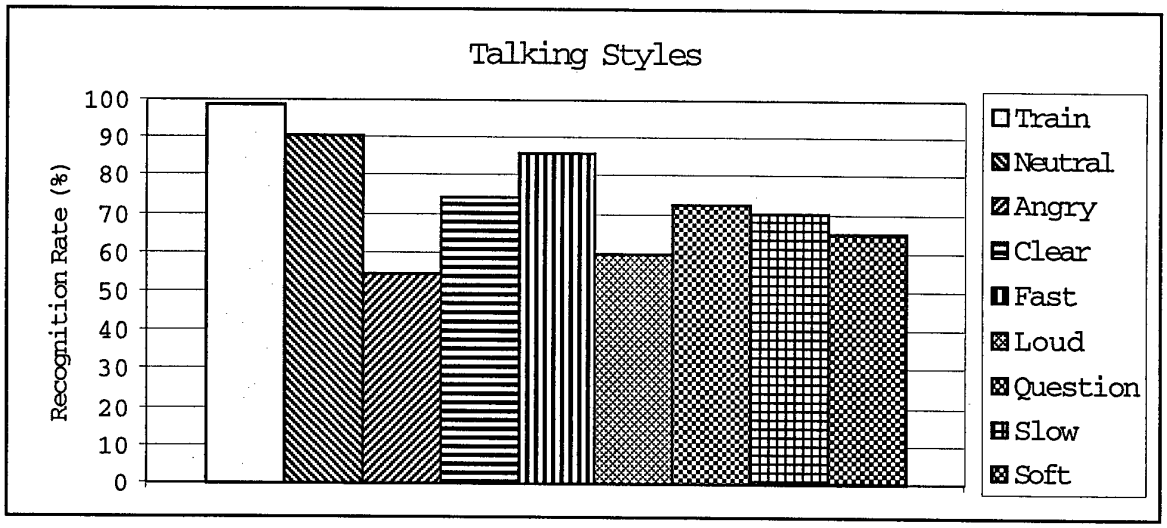


Figure 6.7: Average recognition rate in the simulated stress conditions.

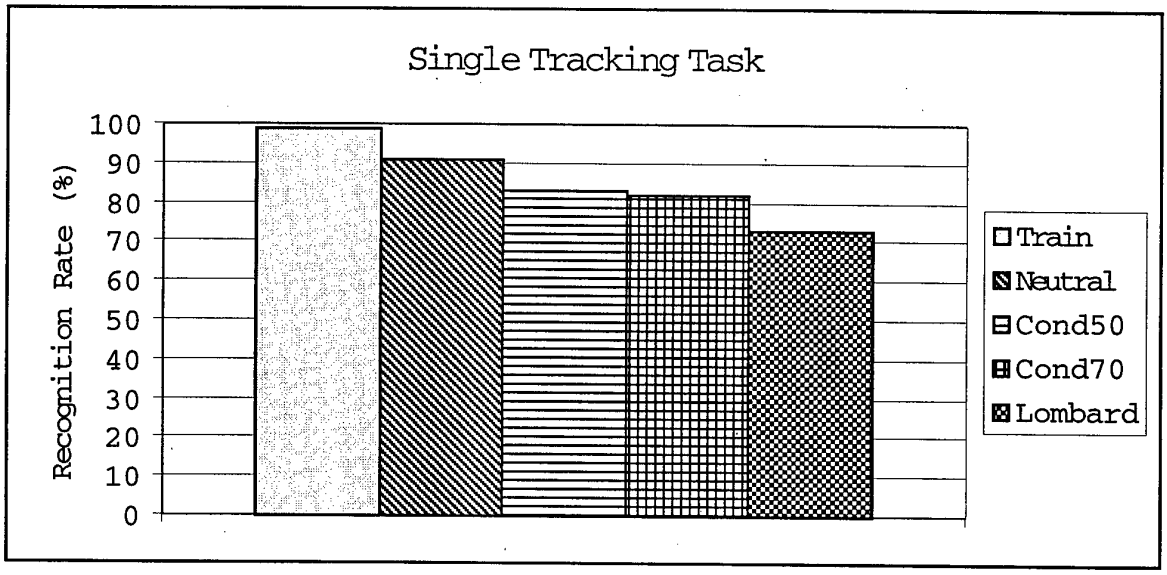


Figure 6.8: Average recognition rate in the simulated task domains and Lombard effect stress conditions

Discussion: The ASR system used in these experiments was not explicitly optimized for stress conditions, and therefore these results can be used as a reference for further studies. The results show the following tendencies:

- Recognition performance varies considerably across the different talking styles and stress conditions.

- There was no significant difference between the recognition rate under moderated stress (cond50) and high tracking workload stress (cond70).

In this simulated stress domain, we observe that neutral speech had the best recognition rate. For the talking-style domain, we observe that the results for most of these experiments was not significantly different. The talking styles can be grouped in three classes. All the members of each class have overlapped reliability intervals.

- Class A Neutral and fast.
- Class B Clear, question, soft, slow and loud.
- Class C Angry and loud.

Class C had the worst recognition performance, and Class A the best.

For the single tracking task domain, the results for moderated stress (cond50) and for high stress (cond70) do not present significant differences.

6.2.2.4 Test conducted by RSPL

The discussion in this section is based on evaluations and improved algorithm formulation for speech recognition under stress from the Robust Speech Processing Lab (RSPL), Univ. of Colorado/Duke Univ.

To illustrate the problem of speech recognition in stress and noise, a baseline speech recognizer (VQ-HMM) was employed on noisefree and noisy stressed speech from SUSAS. This system was a discrete observation, 5-state, left-to-right HMM, trained in a speaker dependent mode (trained using odd neutral tokens; and tested using even tokens then repeated with even training, and odd token testing). A 64-state speaker-dependent VQ codebook was trained using two minutes of training data. Figure 6.9 shows that when stress and noise are introduced, recognition rates decrease significantly. When white Gaussian noise is introduced, noisy stressed speech rates varied, with an average rate across all ten stressed conditions of $Avg = 30.3\,\%$ (i.e., a 58 % decrease from the 88.3 % neutral rate). Recognition performance also varies considerably across the ten stressed speaking conditions as reflected in the large standard deviation in recognition ($SD = 15.35$ for noise free and 9.12 for noisy stressed conditions).

6.2.2.5 Speaker-independent task-independent continuous recognition system

Experiments conducted by the DERA Speech Research Unit (SRU) investigated the effect of speech under stress on an automatic speech recognition system which had specifically not been optimized for either the speakers or the vocabulary under test. A recognition system was configured which was 'speaker-independent'—that is, where the speech models had been generated from speakers outside the SUSAS data set—and task-independent—that is, where the word models had been constructed from general un-stressed speech material. The speech models used were a set of well-trained context-dependent sub-word HMM models which had previously been optimized by the SRU for large vocabulary continuous speech recognition over the telephone.

The original speech training data had been derived from full-bandwidth Wall Street Journal (WSJ-SI284) recordings. The resulting models had then been adapted to telephone bandwidth speech. These models had the same recording bandwidth as SUSAS but, because of their telephone characteristics, the SUSAS input data needed to be 'normalized' prior to recognition. Normalization was applied by first performing unconstrained phone recognition (that is, phones were recognized with no restriction on their sequential ordering), and then by using 'spectral shape adjustment' (SSA) to adjust the speech vectors accordingly.

Since the SUSAS data consists of words spoken in isolation, the SRU continuous recognition system was configured to employ a single-word syntax. This limited the opportunity for word insertion errors, and allowed the results to be more directly comparable with those arising from

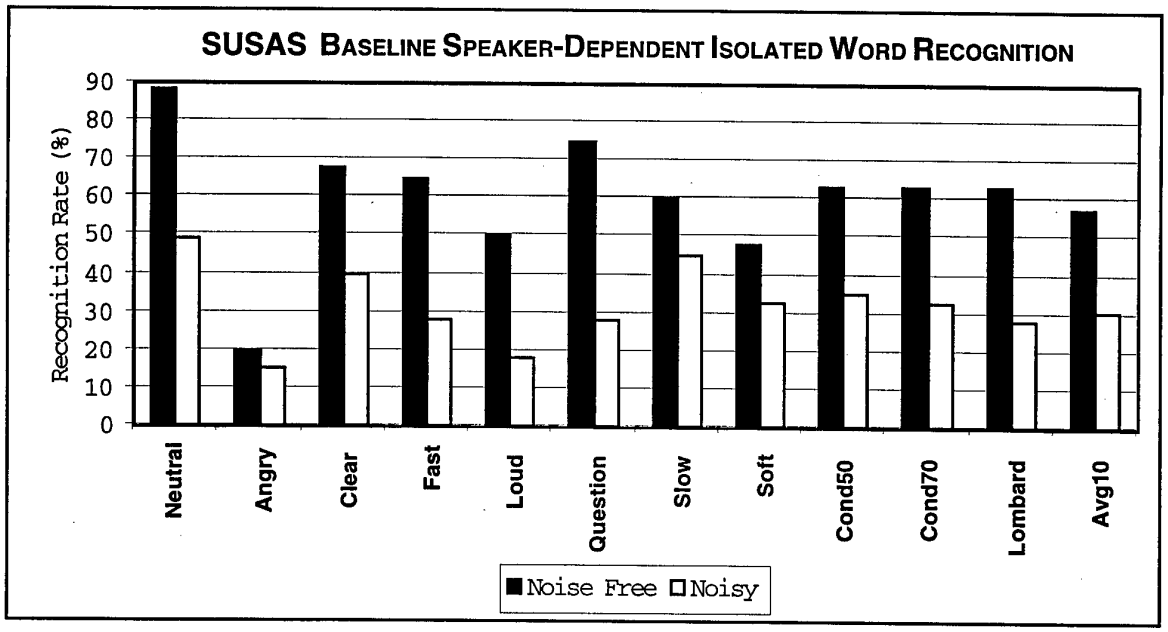


Figure 6.9: Recognition performance of neutral and SUSAS stressed type speech in noise free and noisy conditions. (Additive white Gaussian noise, SNR = +30 dB)

experiments performed using isolated word recognition. The vocabulary was defined to be the 35-words in the SUSAS data set, and each word was allocated a single phonetic pronunciation.

The results for the eleven 'simulated' stress test conditions are illustrated in Figure 6.10. Overall, the average word recognition rate was around 77 %. The worst condition was 'angry' with a word recognition rate of around 54 %. The 'neutral' condition gave rise to a word recognition rate of 86 %, but the best condition was 'Cond50' at 87 % of words correct.

The results for the four 'actual' stress test conditions are illustrated in Figure 6.11. In this case the overall average word recognition rate was 41 %—that is, about half that achieved under the simulated stress conditions. However, the results were distinctly bi-modal with the 'SM' condition being the worst at 30 % word recognition rate together with the 'FF' condition giving rise to 44 % of word correct. The other two conditions, 'Medst' and 'Hist' both had word recognition rates around 81 %, which is comparable with several of the 'simulated' stress conditions.

These results indicate that, whilst useful performance (that is, around 90 % word recognition rate) is just achievable from an unadapted state-of-the-art task-independent speaker-independent system in benign conditions, performance deteriorates badly for most stressed conditions.

6.2.2.6 Large vocabulary continuous speech recognition system

The DERA SRU were able to use the recognition framework outlined in the previous section to perform large vocabulary recognition on the SUSAS data. The same model set was used but, in this case, the vocabulary was not limited to the 35 words in the SUSAS data, but expanded to include the 20 000 most frequent words in the Wall Street Journal Training corpus (the 20 000 words being themselves derived from a set of 200 000 words). SSA was again used to compensate for the different spectral characteristics of the WSJ and SUSAS data sets. Similarly, a single-word syntax was used to limit insertion errors.

The results for the eleven 'simulated' stress test conditions are illustrated in Figure 6.10. Overall the average word recognition rate was around 29 %—that is, about one third of the

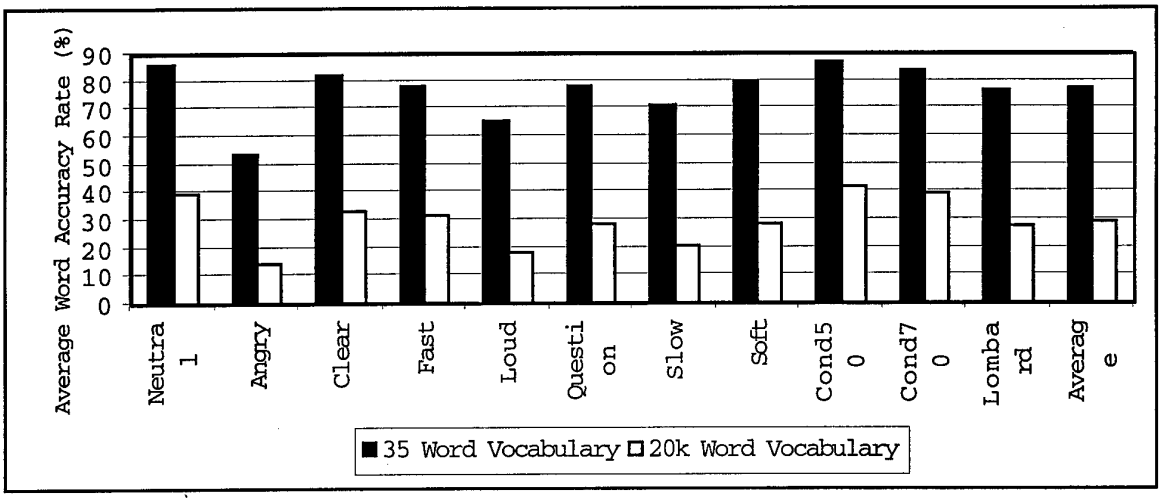


Figure 6.10: Recognition performance of two speaker-independent continuous speech recognition systems on the SUSAS simulated stress database.

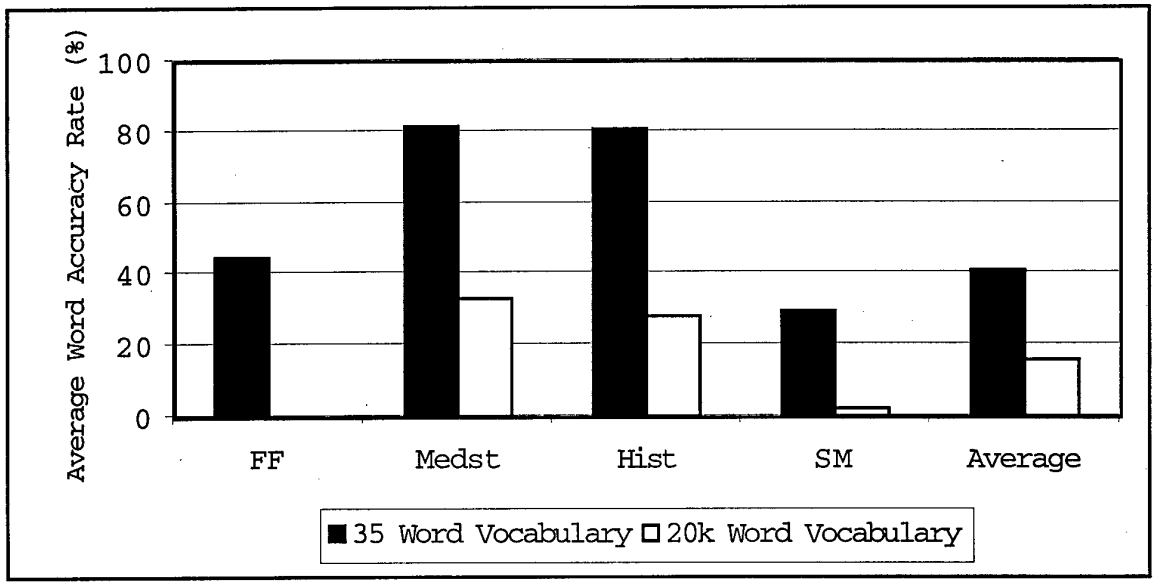


Figure 6.11: Recognition performance of two speaker-independent continuous speech recognition systems on the SUSAS actual stress database. (Note: the word accuracy rate for the 20k vocabulary system on the FF data was 0 %.)

recognition rate for the 35-word system. The worst condition was again 'angry' with a word recognition rate of around 14 %. The 'neutral' condition gave rise to a word recognition rate of 40 %, but the best condition was again 'Cond50' at 42 % of words correct.

The results for the four 'actual' stress test conditions are illustrated in Figure 6.11. In this case the overall average word recognition rate was 16 %. As in the previous section, the results were bi-modal with the 'SM' condition being the worst at 2 % word recognition rate and the 'FF' condition with 0 % of words correct. The other two conditions, 'Medst' and 'Hist' both had word recognition rates around 30 %, which is comparable with the average 'simulated' stress conditions.

Clearly, these results indicate that an unadapted state-of-the-art LVCSR system is unable to achieve any degree of useful performance in any of the stressed speech conditions. Thus it is concluded that system optimization/adaptation (either to the speaker, to the task or to the stress conditions) is of paramount importance to achieving a usable level of performance.

6.2.2.7 COTS large vocabulary continuous speech recognition system

An evaluation was performed by RSPL on the effects of stress on a commercial large-vocabulary continuous speech recognition system. The purpose here is to demonstrate that the effects of stress have a serious impact by degrading recognition performance for commercial systems. Speech data from the SUSAS corpus was used for the evaluation. Since SUSAS data represents mostly isolated words, a series of carrier phrases was used to submit stressed speech tokens to the commercial system. All recognition results reflect only the performance of the system on the stressed speech, since the carrier phrases were held constant across the stressed speaking conditions.

6.2.2.8 RSPL ViaVoice Gold Experimental Set-up

IBM's VIA-VOICE GOLD is a commercial large vocabulary continuous speech recognition (LVCSR) system. It has a 64,000 word vocabulary which can be extented to 250,000. The performance of this system can be improved by enrolling the user with an initial training session where the HMM recognition models are adapted to the speaker, however, this was not carried out since the focus of this experiment was on the performance of the use of the system as a commercial off-the-shelf recognizer, without attempting to improve it's performance by any means. This system is believed to have a bigram or a trigram language model which renders the recognition of isolated words difficult if not impossible. In order to by-pass this problem SUSAS words were presented to the system within grammatically correct sentence structures where each sentence was composed of a leading carrier phrase followed by the SUSAS words. The SUSAS vocabulary set consists of 35 words, whereas only 16 different carrier phrases were shared among the input word-list. The list of words as well as carrier phrases are summarized in Table 6.2.

Each SUSAS word was concatenated with the respective carrier phrase and recorded onto a DAT tape player/recorder. In order to isolate each sentence from the following example a command phrase, *period new line*, was appended to each carrier phrase. This ensures that each sentence was completed with a period and the cursor was advanced to the next line. The tape was played to the IBM recognizer through the line-in input of the computer. The output of the recognizer, the recognized text stream, was printed into a text file. This text file was then processed to evaluate the system performance. A block diagram describing the experimental set-up is shown in Figure 6.12.

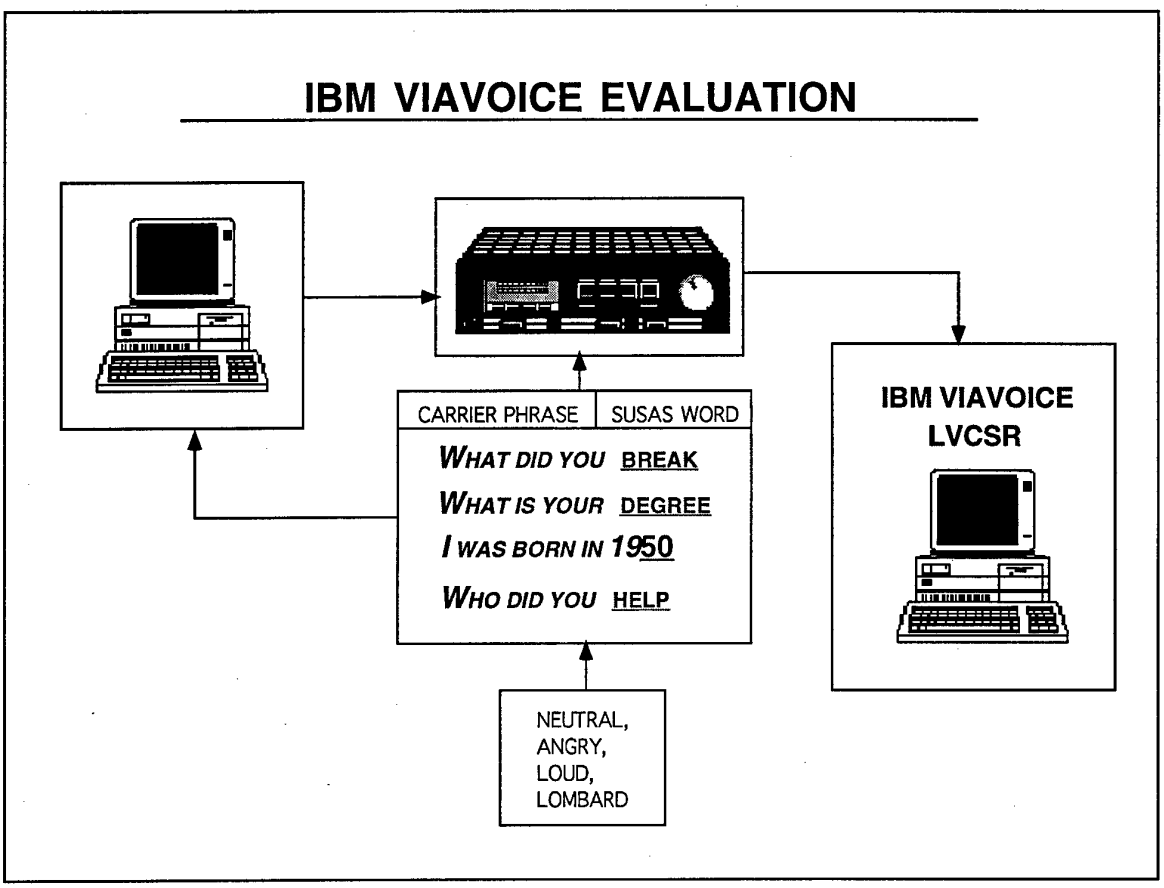


Figure 6.12: Block diagram of the IBM Via-Voice Gold large-vocabulary continuous speech recognition evaluation using the SUSAS speech under stress database. Isolated speech under stress tokens were automatically inserted into carrier phrases for LVCSR testing.

Table 6.2: SUSAS Words and Carrier Phrases

What did you *break*	What is *nav*
What did you *change*	I said *no*
What is your *degree*	My phone number is 684 354 *oh*
What is your *destination*	What did you put *on*
I will go *east*	When did you go *out*
My phone number is 684 354 *eight*	What is your *point*
I was born in 19 *eighty*	My phone number is 684 354 *six*
Where did you *enter*	I will go *south*
I was born in 19 *fifty*	Where did you *stand*
What did you *fix*	How did you *steer*
Why does water *freeze*	How did you *strafe*
What did you *gain*	I was born in 19 *ten*
Where did you *go*	I was born in 19 *thirty*
I said *hello*	My phone number is 684 354 *three*
Who did you *help*	His car is *white*
I plotted a *histogram*	His car is *wide*
The water was *hot*	My phone number is 684 354 *zero*
How did you *mark*	

6.2.2.9 RSPL ViaVoice Gold Evaluations

The following three SUSAS simulated stress styles, *Angry, Loud, Lombard*, and the SUSAS *Actual* stressed speech domain, in addition to *Neutral* speech, were included in the analysis. Two methods were used to assess performance of the system. First, errors in the recognized carrier phrase were ignored and only correctly recognized SUSAS words were counted. Homonym type outputs were also counted as errors (i.e., *histogram* and *his stick gram*). The results, based on this type of evaluation method, are shown in Figure 6.13. The second evaluation method was based on the standard LVCSR NIST scoring method at the word and sentence levels. The results for the second method are given in Figure 6.14.

For the first evaluation method, the neutral spoken SUSAS words achieved the highest recognition rates (38.6%), followed by Lombard (32.7%), loud (26.2%), angry (23.2%), and actual (0.0%). After considering the selected carrier phrases and the SUSAS words presented, the overall results for each of the stress conditions were revised. One example of carrier phrase impacting the recognition evaluation involved numerals (e.g., "The telephone number ... three eight *four oh*," is recognized as *for all*). On the other hand, it is believed that some of the SUSAS words are not in the dictionary (i.e. *nav, strafe, histogram* (or their probability is so low that when they are recognized, the language model drops the word and selects an alternative word with higher overall probability). The revised results were obtained by excluding these factors (revised overall column in Figure 6.13). While these factors do increase the recognition rate, the order of results across the four stress conditions remain the same.

In the NIST evaluations, two set of results are reported, one corresponding to sentence level, the other to word level accuracy. While there are a total of 630 SUSAS words for Simulated Stress Conditions (*Neutral, Angry, Loud, Lombard*), there are 484 words for *Actual* stress conditions. The number of words in the entire sentences (including carrier phrase) is 3276 for the Simulated stress set, and 2486 for the actual stress sentences. The rank order of the stress conditions follows the same pattern as the first evaluation method. The reason for the similar word accuracy results was due to the ratio of SUSAS words in the test set to the overall number of words in the sentences (19.2%).

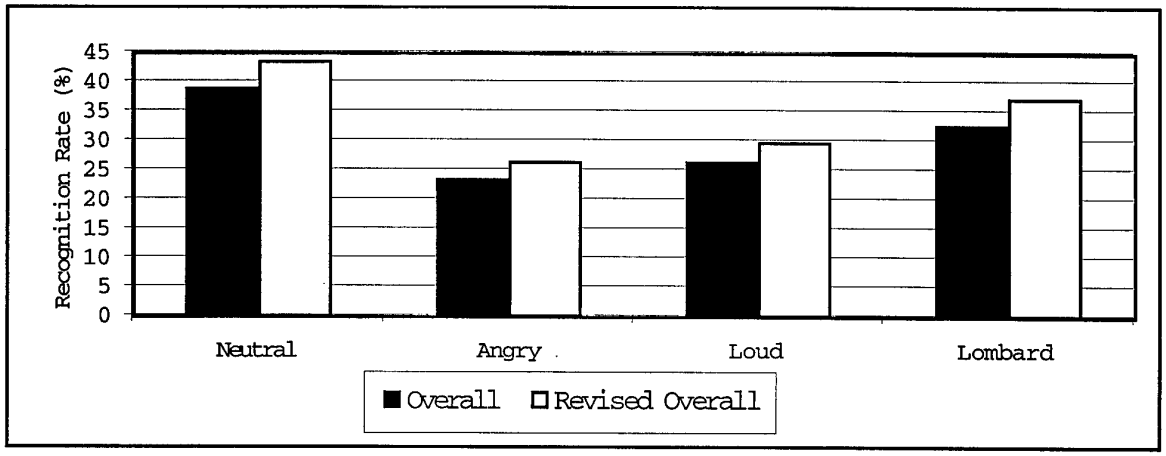


Figure 6.13: Recognition Results for IBM ViaVoice Gold Using SUSAS Database.

It can be clearly seen that in a commercial LVCSR framework, speech under stress impacts overall recognition performance beyond what is expected for neutral speech. The reader should note that there are a number of reasons for obtaining low IBM ViaVoice recognition rates for *Neutral* speech. Some of these include: (i) not having an initial enrollment training session to

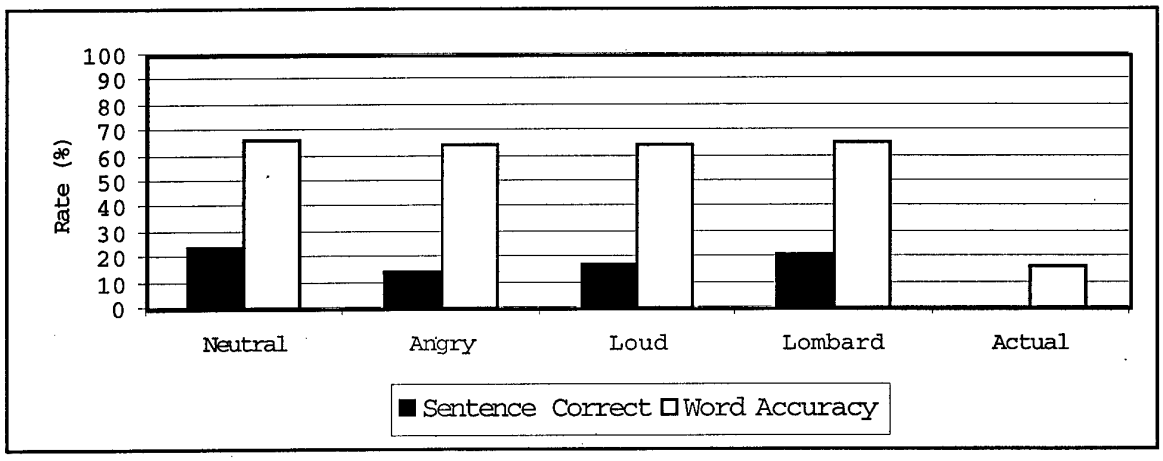


Figure 6.14: Recognition Results for IBM ViaVoice Gold Using SUSAS Database. (Note: the sentence correct rate on the actual data was 0%.)

adapt the models to the speaker[1], (ii) the effects of inserting SUSAS words within a carrier sentence, and (iii) the selection of the specific carrier phrases. What is important to note is the relative difference of the recognition rate for the *Neutral* reference speech and the recognition rates for the different stressed speaking conditions reflecting the impact of stress on recognition performance.

6.2.2.10 Discussion

Figure 6.15 shows a summary of all the experiments carried out using the SUSAS database. It can be seen from the figure that all systems had more trouble recognizing certain stressed speech conditions than others. For example, angry speech had the worst recognition rate for all systems tested in these experiments. If the speech condition was considered in the development of the system then there is not too much degradation in the recognition rate. For example, recognition systems handle differences in the rate that the speech is spoken and most systems did not see too much degradation when comparing the neutral to the fast conditions, although the systems did not do as well with the slow condition. The system evaluations using speech under stress from the actual domains of SUSAS (e.g., roller-coaster rides, military helicopter pilot speech) showed a more significant reduction in recognition performance (i.e., a reduction from 96% to 27%).

The results show that systems with a vocabulary limited to the words in the database performed better than those with a large vocabulary. Comparing the DERA 35 word vocabulary and the RSPL and GTH isolated word systems with the DERA 20K vocabulary and the ViaVoice system results we see that the small vocabulary systems have better recognition rates under all speaking conditions. It is interesting to note that while the baseline speaker dependent systems used by RSPL and GTH gave slightly better results with neutral speech when compared to the DERA LVCSR, they performed worse than the LVCSR system on some of the speaking conditions. This is probably due to the difference in available training data between the speaker dependent systems (using data from SUSAS), versus LVCSR (which was already trained using several orders of magnitude more data from a wider speaker population). The evaluations performed by RSPL on the Viavoice LVCSR system required that artifical conditions be imposed by inserting the words in a carrier phrase. It is also possible that continuous speech is a better

[1]This was not performed, because the carrier phrases were produced by a speaker outside of the SUSAS speaker set, so adapting the recognizer could not be achieved for both the carrier phrase test speaker and SUSAS test speaker at the same time

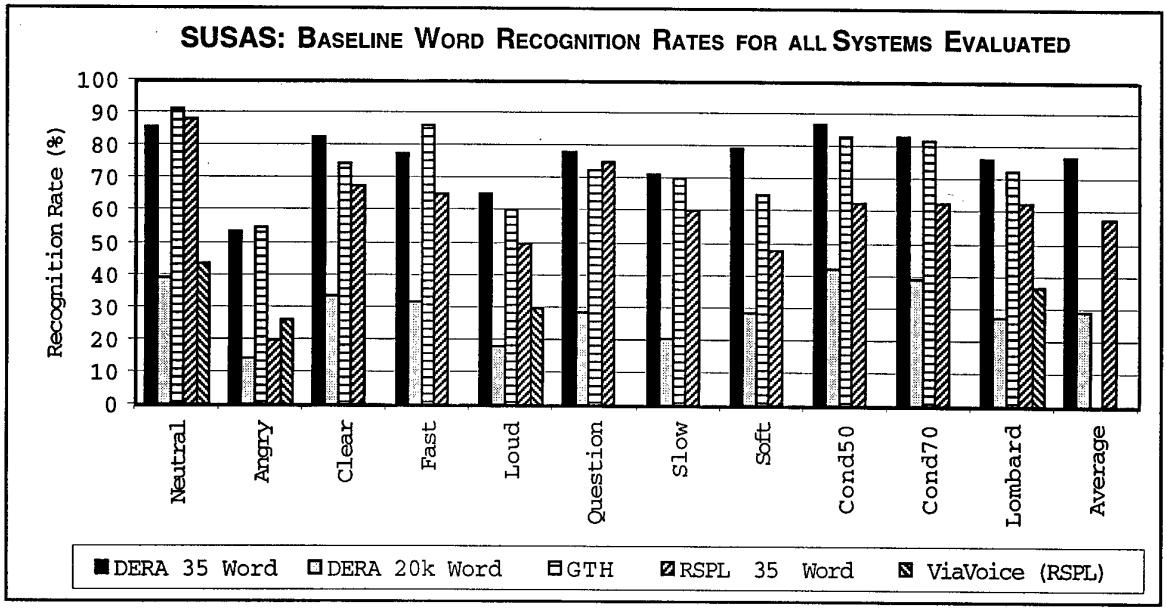


Figure 6.15: Word Recognition Rate for All tests Conducted on the SUSAS Database. Note that some experiments were run on subsets of the database.

representation of the speech samples than is expected by the GTH isolated recognition system. It is very important, however, to keep in mind that a direct comparison of the recognition systems in Figure 6.15 is done only for reference, since the LVCSR systems were previously trained with significant amounts of data not made available to the speaker-dependent systems.

6.2.3 Stress Compensation Techniques

6.2.3.1 Background of Recent Methods for Stressed Speech Recognition

Approaches for robust recognition can be summarized under three areas: (i) better training methods, (ii) improved front-end processing, and (iii) improved back-end processing or robust recognition measures. These approaches have been used to improve recognition of speech in noisy and Lombard effect environments, as well as workload task stress or speaker stress conditions.

To formulate automatic speech recognition algorithms which are more effective in changing environmental conditions, it is important to understand the acoustic-phonetic differences between normal speech and speech produced under stressed conditions (some of these issues were considered in Chapter 4). Several studies have shown distinctive differences in phonetic features between normal and Lombard speech [53, 54, 84, 146], and speech spoken in noise [51]. Further studies have focused on variation in speech production brought on by task stress or emotion [14, 54, 57, 68, 71]. The primary purpose of these studies has been to improve the performance of recognition algorithms in Lombard effect [84, 74, 148], stressed speaking styles [96, 123, 30], noisy Lombard effect [54, 66, 58], and noisy stressful speaking conditions [131, 54, 71, 57].

Approaches based on improved training methods include multi-style training [96, 123], simulated stress token generation [14, 64], and others such as training/testing in noise and robust distance measure approaches [41, 83]. Improved training methods can increase recognition performance, however results degrade as test conditions drift from the original training data. In fact, even if background noise could be addressed in this manner, poor recognition performance will persist due to changing speech characteristics caused by stress and Lombard effect. Further discussion of alternative methods for stressed speech recognition which include front-end

processing/speech feature-estimation can be found in the following references [60, 85].

6.2.3.2 Stress Compensation Methods for Speech Recognition

Since noise, stress, and Lombard effect have been shown to disrupt speech recognition, researchers at RSPL have considered alternative speech modeling/feature processing methods based on a *Source Generator Framework*. The notion of Source generator representation was considered in a study by Hansen, 1994 [58], and later employed in other robust recognition algorithms [57, 14, 24, 68, 64].

The concept of a source generator framework is as follows. For the production of a word, it is assumed that a sequence of coordinated movements of the vocal system articulators and excitation controls are needed (represented in the multi-dimensional speech production space). The coordinated sequence of excitation and articulatory controls are modeled as a smooth path in this speech production space. It is hypothesized that vocal system controls (i.e., articulators, etc.) will be perturbed under stressed speaking conditions resulting in deviations from this "neutral" production space path. From previous studies, it is known that the presence of stress will cause changes in phoneme production with respect to glottal source factors, pitch, intensity, duration, and spectral shape [54, 70]. The framework suggests that the perturbation of these vocal system controls can be modeled by a change in the speech source generator γ_j in some F-dimensional feature space. Each source generator will occupy some volume in the multi-dimensional feature space, and that deviations in speech production under stress will result in a feature sequence which deviates from the mean "neutral" path. A more complete discussion can be found in [60].

Three front-end processing approaches are proposed to compensate for the changes due to stress for speech recognition. The formulated methods are based on speech parameter estimation schemes which are less sensitive to varying levels and types of background noise, as well as accurate modeling of the human speech production under stress to improve recognition in adverse environments. These methods employ robust speech feature estimation algorithms, as well as stress equalization techniques based on source generator theory. A comparison of how the stress equalization methods are applied to the extracted speech feature sequence is shown in Figure 6.16.

6.2.3.3 Combined Stress Equalization & Noise Suppression

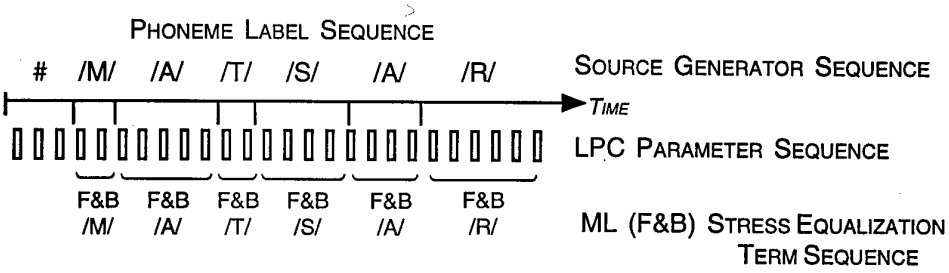
The first front-end approach employs feature enhancement and production equalization algorithms under the source generator framework [60, 72]. The intent here, was to demonstrate that a direct stress equalization can reduce the effects of stress for robust recognition in diverse environmental conditions. Therefore, though the choice of source generator type is arbitrary, hand labeled phoneme partitions were employed (see Figure 6.16a). The feature enhancement algorithm was formulated based on a class of constrained iterative techniques previously derived for automatic enhancement of speech in varying background noise environments. The enhancement technique employs speech specific inter and intra-frame spectral constraints applied to line-spectral-pair parameters and autocorrelation estimates [69]. Next, a multi-dimensional stress equalization approach was formulated which produces recognition features which were suggested to be less sensitive to varying factors caused by stress. The stressed based equalization domain was restricted to be the spectral domain in an eight-dimensional feature space ($\vec{k} = d_1 \cdots d_8$).

$$\Psi_{\text{SPECTRAL}(\vec{k}),c}[\gamma_j] : s_{\gamma_j}^{\text{neutral}} \mapsto s_{\gamma_j}^{\text{stressed}} \tag{6.3}$$

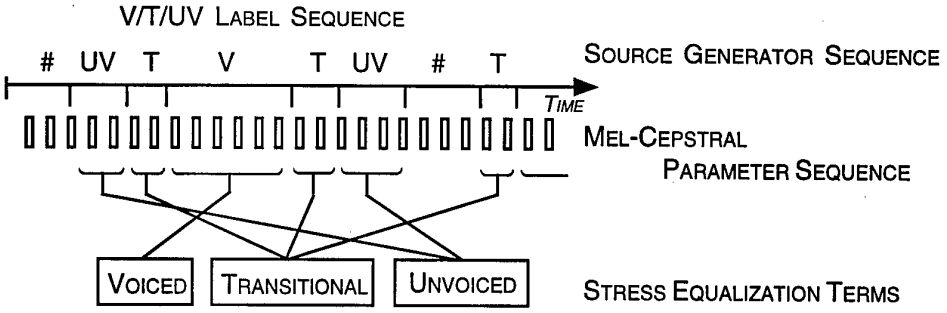
$c \in$ (NEUTRAL, SLOW, FAST, SOFT, LOUD, ANGRY, CLEAR, QUESTION, C50 TASK, C70 TASK, LOMBARD)

The spectral dimensions SPECTRAL$\vec{k}$ were defined as the first four formant locations and bandwidths $(\vec{F}, \vec{B})$. Stress equalization of the speech feature set was achieved using a unique trans-

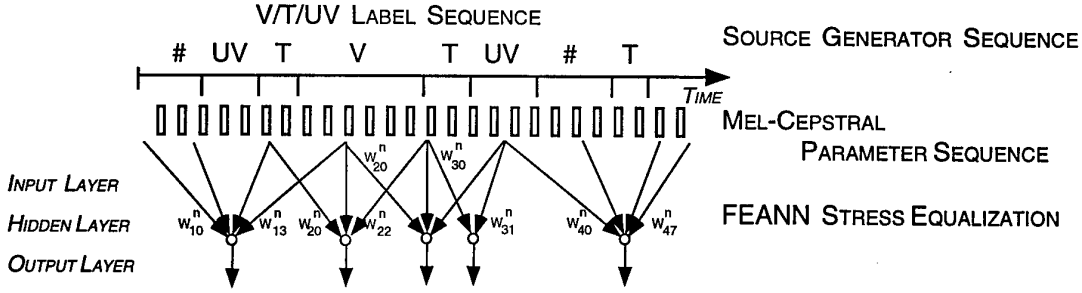
(a.) **(F_N & B_N) STRESS EQUALIZATION & (AUTO:I,LSP:T)**



(b.) **FIXED (VOICED/TRANSITIONAL/UNVOICED) ML STRESS EQUALIZATION**



(c.) **DEPENDENT FEANN (VOICED/TRANSITIONAL/UNVOICED) ML STRESS EQUALIZATION**



(d.) **SEQUENCE ADAPTIVE (VOICED/TRANSITIONAL/UNVOICED) ML STRESS EQUALIZATION**

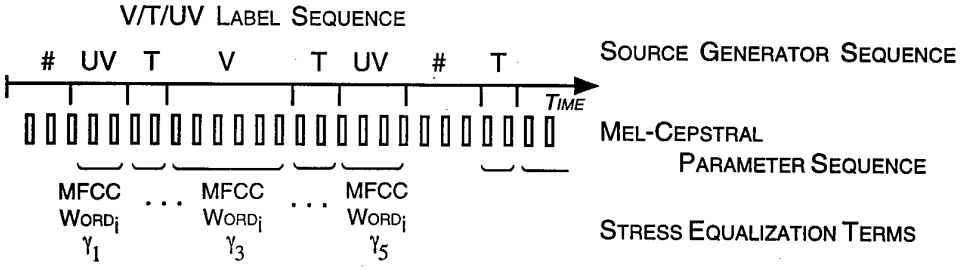


Figure 6.16: A comparison of how stress equalization is applied to extracted speech features. (a.) Stress equalization of formant location and bandwidth across a source generator phoneme sequence. (b.) Stress equalization of Mel-cepstral features using fixed compensation terms across a V/T/UV source generator sequence. (c.) Stress equalization of Mel-cepstral features using a word dependent feature enhancing artificial neural network (FEANN) across a V/T/UV source generator sequence. (d.) Stress equalization of Mel-cepstral features using word dependent Maximum-Likelihood compensation across a V/T/UV source generator sequence.

formation term $\mu_{\text{SPECTRAL}}(\Psi[\lambda, d_i, \gamma_j])$, which was estimated for each feature dimension d_i, stressed condition $c = \lambda$, and source generator γ_j as

$$\mu_{\text{SPECTRAL}}(\Psi[\lambda, d_i, \gamma_j]) = \frac{\frac{1}{N_j} \sum_{t_n=t_1}^{t_{N_j}} \psi_{i,j}(t_n)}{\frac{1}{N_{j,\lambda}} \sum_{t_n=t_{1,\lambda}}^{t_{N_{j,\lambda}}} \psi_{i,j}^{(\lambda)}(t_n)}. \tag{6.4}$$

Next, using a hidden Markov model recognition framework, baseline scores were obtained (Figure 6.17) for SUSAS speech under neutral, stressful, noisy neutral, and ten noisy stressful speaking conditions (e.g., loud, angry, computer task conditions, Lombard effect, etc.). Combined stress equalization with constrained feature enhancement was shown to reduce the average word error rate for recognition of noisy stressful speech by $-38.7\,\%$ (mean recognition for noisy stressful speech increased from $30.3\,\%$ to $57.3\,\%$). Significant improvement occurred for noisy speech under loud, angry, and Lombard effect stress conditions. The tandem recognition algorithm was also shown to be more consistent across noisy stressful conditions as measured by a decrease in the standard deviation of recognition rate (from 9.1 to 5.7). Further details can be found in previous studies (Hansen, et al., 1988,89,95 [54, 71, 72]). The results suggest that the combination of a flexible source generator framework to address stressed speaking conditions, and a feature enhancement algorithm which adapts based on speech specific constraints, can be effective in reducing the consequences of stress and noise for robust automatic recognition.

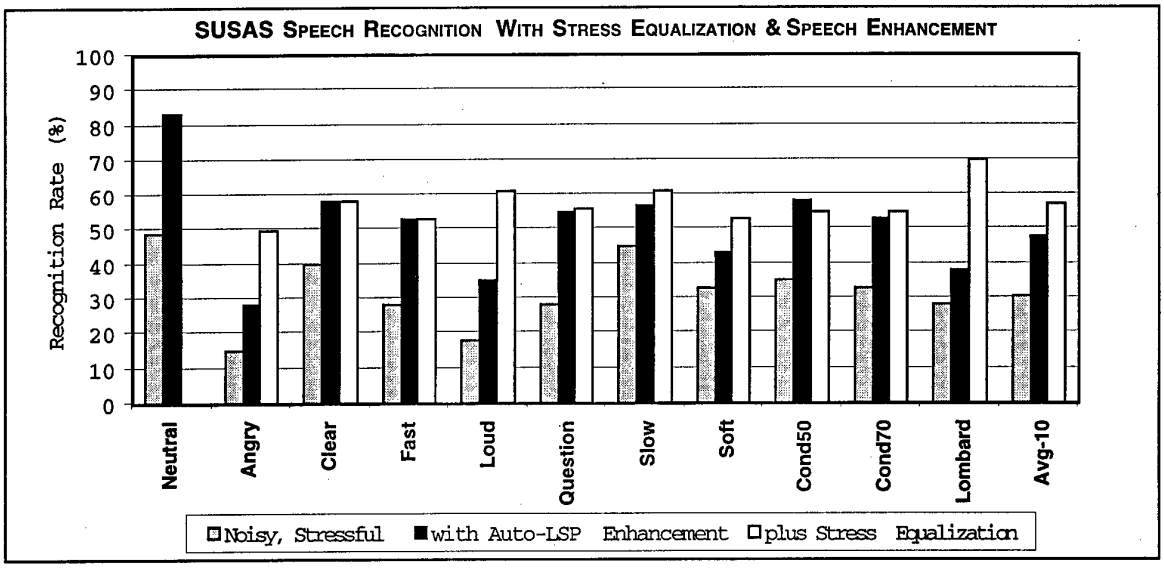


Figure 6.17: Recognition performance of noisy stressful speech with combined generator enhancement and stress equalization (Auto is with Auto:I,LSP:T front-end speech enhancement and Equalization includes $(\vec{F}), (\vec{B}), (\vec{F}\&\vec{B})$ (formant location/bandwidth) Stress Equalization. Note: Additive white Gaussian noise, $+30\,\text{dB}$ SNR

6.2.3.4 Fixed ML and FEANN stress equalization

While useful, the maximum likelihood (ML) formant location and bandwidth based stress equalization method requires phoneme level sequence information. Front-end modifications have also been proposed which normalize the spectral characteristics of stress speech so that stress speech parameters resemble neutral speech [30, 66, 58]. In the compensation method by Chen [30], the impact of stress was assumed to remain constant across an entire word interval, resulting in a fixed whole word compensation stress vector. From Chapter 4, it was shown that the impact of

stress was not uniform over an entire sequence of phonemes, so compensation at the sub-word level should be more effective.

The next set of front-end stress equalization methods, developed by RSPL, removes the requirement of phoneme level sequence information. The second approach is based on a maximum likelihood stress equalization method which normalizes input speech feature sequences using a set of fixed equalization terms (see Figure 6.16b) [66]. This method assumes that input speech is parsed into a sequence of voiced/transitional/unvoiced (V/T/UV) labeled sections [56], and that three different maximum likelihood stress equalization vectors for voiced, transitional, and unvoiced speech sections are employed to compensate for the effects of stress. Results using SUSAS speech data show that stress compensation using three fixed V/T/UV stress equalization terms improves Lombard speech recognition performance by +10 %. This method was later adapted for real-time implementation and evaluated for ten noisy stressful conditions with a +17 % improvement in recognition [68, 24].

Another approach using a feature enhancing artificial neural network (FEANN) was also developed by RSPL which reduces stress effects during parameterization [31, 60]. Figure 6.16c illustrates the basic approach. Here, a unique FEANN was formed for each keyword model and evaluated using a semi-continuous HMM recognizer followed by a likelihood ratio test for keyword detection.

This system was evaluated for the task of keyword spotting of speech under stress using data from the SUSAS database. Results using receiver-operating-characteristic (ROC) curves show that the feature enhancing neural network was able to enhance Mel-frequency cepstral coefficients (MFCC) under stress and reduce the probability of false acceptances of non-keywords by adapting its weights and input layer width based on extracted speech characteristics. Keyword recognition evaluations show that FEANN reduced the number of false acceptances for neutral and Lombard stress by more than one third.

6.2.3.5 MCE-ACC Stress Equalization & Noise Suppression

In this section, robust speech recognition in stress and noise is accomplished via morphological constrained feature enhancement (MCE) and stressed source compensation which is unique for each source generator across a stressed speaking class (see Figure 6.16d)[58] The algorithm developed by RSPL uses a noise adaptive (V/T/UV) boundary detector [56] to obtained a sequence of source generator classes, which is used to direct MCE parameter enhancement and stress compensation. This allows the parameter enhancement and stress compensation schemes to adapt to changing speech generator types. The algorithm is entitled Morphological Constrained feature Enhancement with Adaptive mel-Cepstral Compensation based hidden Markov model recognition (*MCE-ACC-HMM*). The source generator sequence of MCE estimated spectral responses $\hat{\mathbf{S}}_{\gamma_{b_j},\alpha,\beta,\mathcal{D}_g}(\omega_i)$, are then submitted for stress equalization. Stressed speaking conditions are addressed by the choice of a modified source generator for each phoneme-like section. The unknown stress dependent model parameter $C_{\Psi_i(b_j)}(k)$ is estimated by maximizing the log-likelihood function, resulting in the ML estimate. A compensation model vector $\hat{\vec{C}}_{\Psi_i(b_j)}$ is therefore estimated for each detected source generator section during HMM training, and applied during recognition evaluation.

The algorithm was evaluated using SUSAS speech data for noise free and nine noisy Lombard effect speech conditions which include additive white Gaussian, slowly varying computer fan, and aircraft cockpit noise [58] (see Figure 6.18). System performance was compared to a traditional VQ-HMM recognizer with no embellishments. Employing individual recognition scores for all 27 noisy Lombard effect stress conditions, the final mean recognition rate increased from 36.7 % for VQ-HMM to 74.7 % for MCE-ACC (+38 % improvement). The MCE-ACC was also shown to be more consistent, as demonstrated by a decrease in standard deviation of recognition from 21.1 to 11.9, and a reduction in confusable word-pairs. These results demonstrate the consistency of

MCE-ACC recognition improvement for noisy Lombard effect speaking conditions.

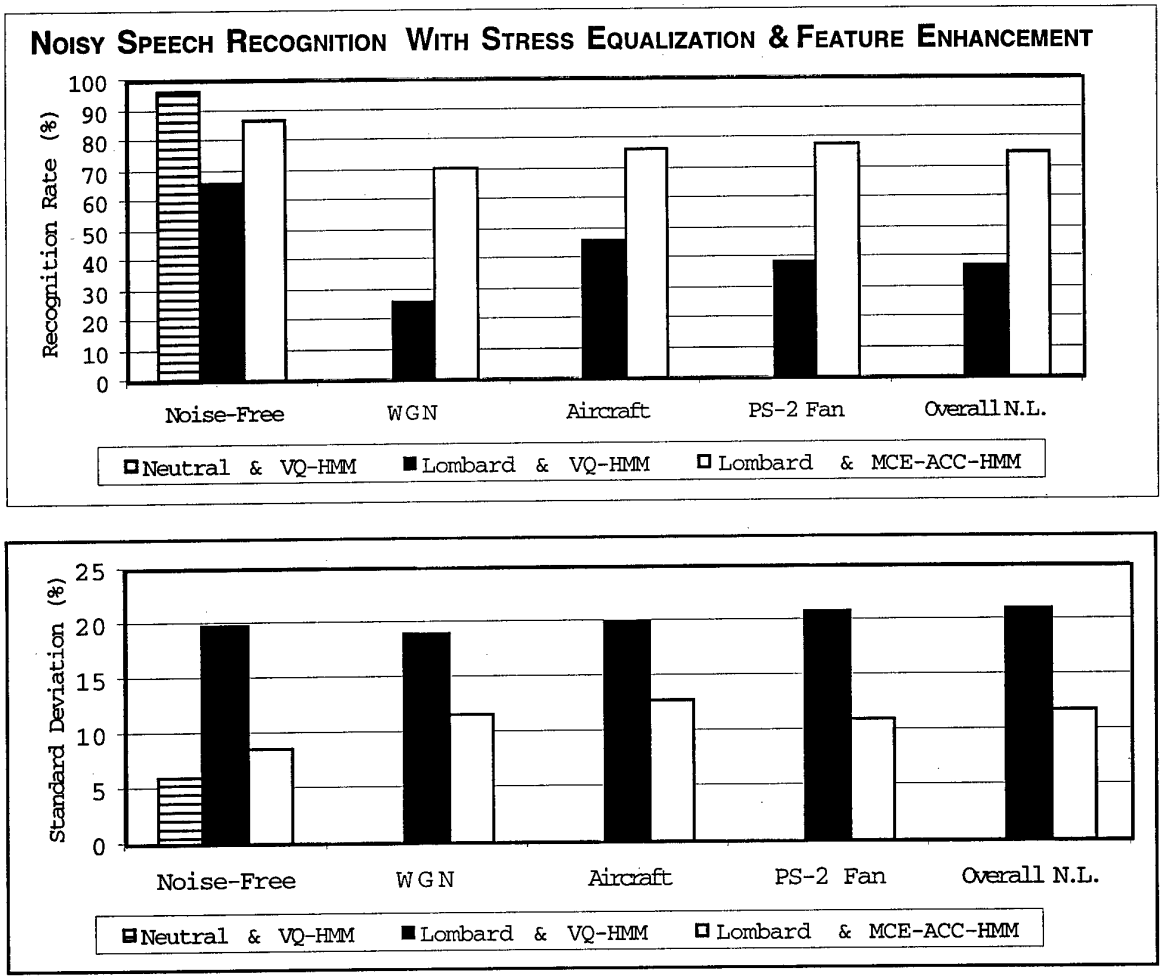


Figure 6.18: Overall recognition results for the VQ-HMM recognizer and the new robust recognizer MCE-ACC-HMM for three types of noise (white Gaussian noise, Lockheed C130 aircraft cockpit noise, IBM PS-2 cooling fan noise). Noise-free and averages over all noisy conditions (10, 20, 30 dB SNR) are shown. The lower plot shows the standard deviation in recognition for each test condition.

6.2.3.6 Stressed Speech Training Methods: Stress Token Generation

A number of studies have been suggested for improving recognition of speech under stress [14, 15, 30, 58, 66, 71, 84, 96], Thus far, the first three approaches to front-end stress compensation outlined in Figure 6.16 have been discussed. While front-end stress equalization can be effective for reducing speech feature variation prior to using a neutral speech trained recognizer, other methods based on system training have been suggested to address stress. Since the performance of a speech recognition system degrades if the recognizer is not trained and tested under similar speaking conditions, an approach called multi-style training by Lippmann *et al.* [96] has been suggested for improving speaker-dependent recognition of stressed speech. This method required speakers to produce speech under simulated stressed speaking conditions and to employ these multi-styles within the training procedure. In addition to significant improvements in stressed speech recognition, this study showed that multi-style training also improved recognition performance under normal conditions. However, a later study by Womack and Hansen [162] showed

that multi-style training actually degrades performance if employed in a limited but speaker-independent application. The cause of this was believed to be due to the issue that speakers have significantly different vocal tract structures. For an individual speaker, multi-style training simply captures the dependent variations a speaker exhibits when asked to speak under the range of simulated stress styles. However, all speakers may not vary the same articulatory structure under stress, and the deviation from the neutral condition for one speaker will not necessarily coincide with the deviation exhibited from neutral for a new speaker.

An alternative technique by Bou-Ghazale and Hansen which also employed the source generator framework turns the stress compensation around by generating simulated stressed tokens which were used for training a stressed speech recognizer [15, 64]. Generating simulated stress data in the training phase rather than compensating for the effect of stress in the recognition phase resulted in a computationally faster recognition algorithm. In the latter approach, both duration and spectral content (i.e., mel-cepstral parameters) were altered to statistically resemble a stressed speech token. Using SUSAS stressed speech data, this method was shown to improve isolated word recognition of Lombard effect speech by 24 % when compared to a neutral trained speech recognizer [64, 14]. This approach therefore extracts a model for how speakers vary speech recognition features from neutral to stress, and applies this model to neutral input training data in order to modify the recognizer word models to address stress. These pertubration word models could therefore also be applied to unseen speakers, since they only reflect the deviation from neutral speaking conditions. A more extensive modeling approach based on HMMs was later proposed for stress modification of neutral speech for synthesis and recognition [18]). This approach will be considered later in Section 6.4 on stressed speech synthesis and coding.

6.2.3.7 Direct Robust Features for Stressed Speech Recognition

The two broad approaches considered in the previous sections for improved stressed speech recognition have focused on (i) front-end stress equalization, where a codebook of stress compensation terms are first estimated and applied to the input feature sequence in some prescribed manner, and (ii) alternative training methods either based on collecting speech data in simulated stress conditions, or formulating models to perturb neutral training tokens to resemble stressed tokens.

The approach taken in this section, is to suggest a direct speech feature set which is reliable for speech recognition in both neutral and stressed conditions. A study by Bou-Ghazale and Hansen [18] at RSPL evaluated the effectiveness of traditional features in recognition of speech under stress and formulated new features which were shown to improve stressed speech recognition. The focus was on formulating robust features which are less dependent on the speaking conditions rather than applying compensation or adaptation techniques. The SUSAS stressed speaking styles considered were angry and loud, Lombard effect speech, and noisy actual (roller-coaster ride) stressed speech. In addition, the study also investigated the immunity of LP and FFT power spectrum to the presence of stress. The results showed that, unlike FFT's immunity to noise, the LP power spectrum was more immune than FFT to stress as well as to a combination of a noisy and stressful environment. Finally, the effect of various parameter processing such as fixed versus variable preemphasis, liftering, and fixed versus cepstral mean normalization were also studied. Two alternative frequency partitioning methods (M-MFCC, ExpoLog) were proposed and compared with traditional MFCC features for stressed speech recognition. The alternate filterbank frequency partitions were found to be more effective for recognition of speech under both simulated and actual stressed conditions. Here, some of the findings from that evaluation are briefly summarized.

Recognizer and Database: The speech data employed was a subset of the SUSAS database. All recognition evaluations were speaker-independent, and considered only male speakers. A 30-

word HMM-based recognizer was formulated using a variable-state, left-to-right model, with 2 continuous mixtures per state. The HMM models were trained with the neutral speech of eight speakers while a ninth speaker was left for open testing. A total of 10 tokens per speaker were employed for training each neutral HMM word model resulting in 80 training tokens per word. The training and testing were done in a round robin scheme to allow all speakers and tokens to be tested in an open evaluation. In evaluating each of the neutral trained HMM models, a total of 2160 tokens were tested from the four speaking styles.

The last evaluation employs actual stressed speech from the SUSAS database. This data consisted of speech produced during the completion of two types of subject motion-fear tasks. The speakers produced speech while participating in two amusement park rides (e.g., a traditional roller-coaster ride and a free-fall ride consisting of a 130 ft vertical drop machine). These two rides were chosen in an attempt to simulate the sudden change in altitude or direction which could be experienced in a military aircraft cockpit under emergency conditions.

Performance of Traditional and Noise-Robust Features in Stress: In this section, the effectiveness of previously proposed noise robust features were investigated in the recognition of stressed speech. The first set of features evaluated in this study were the one-sided autocorrelation linear prediction coefficients (OSALPC) [78]. The OSALPC technique is based on the application of the windowed autocorrelation method of linear prediction to the one-sided autocorrelation sequence as discussed in [77, 78]. OSALPC has been shown to outperform linear prediction (LPC) as well as two other noise robust methods. The second set of features evaluated were the cepstral-based OSALPC, referred to as OSALPCC, which were compared to the performance of traditional cepstral-based LPC and mel-frequency scale coefficients (MFCC). Therefore, the recognition performance of the following set of features was compared (i) linear prediction coefficients (LPC), (ii) linear prediction cepstral coefficients (LPCC), (iii) one-sided autocorrelation linear prediction coefficients (OSALPC), (iv) OSALPC-based cepstral coefficients (OSALPCC), and (vi) mel-scale filter bank cepstral parameters (MFCC).

Next, the performance of these features was considered for recognition of speech under stress. Two sets of evaluations are presented. The first compares the performance of HMM models trained with static features with no parameter processing while the second compares the performance of models trained with static and dynamic features in addition to parameter processing such as cepstral liftering and cepstral mean normalization (CMN). In both evaluations, the models were speaker-independent neutral trained and were tested with SUSAS speech from four speaking conditions : neutral, angry, loud, and Lombard effect.

The results, plotted in Figures 6.19 (static) and 6.20 (static, and delta with parameter processing), show that for both evaluations the one-sided autocorrelation linear prediction coefficients (OSALPC) performed better than traditional LPC for all three stress conditions. OSALPC, however, does not achieve the highest performance among the evaluated features. In fact, the three remaining cepstral features, LPCC, OSALPCC and MFCC, achieved higher recognition rates than OSALPC. In addition, the results show that cepstral based OSALPC outperformed OSALPC by 12.4 % across the four speaking conditions for static features, and by 21.7 % for static and dynamic feature trained models.

MFCC and LPCC parameters achieved the highest recognition rates in both static and combined (static, dynamic, with parameter processing) evaluations. Their performance was very similar across all four speaking conditions as shown in Figures 6.19 and 6.20. Both features achieve a higher level of recognition performance than OSALPCC across all four speaking styles in both scenarios. In summary, these results show that while noise-robust features such as OSALPC may be robust in noise, they are not necessarily robust to the presence of speaker stress. These results also suggest that features derived from cepstral analysis clearly outperform features derived from a linear predictive model. It is recommended that due to the variability across the three stress conditions, new feature sets are needed for improved stressed speech

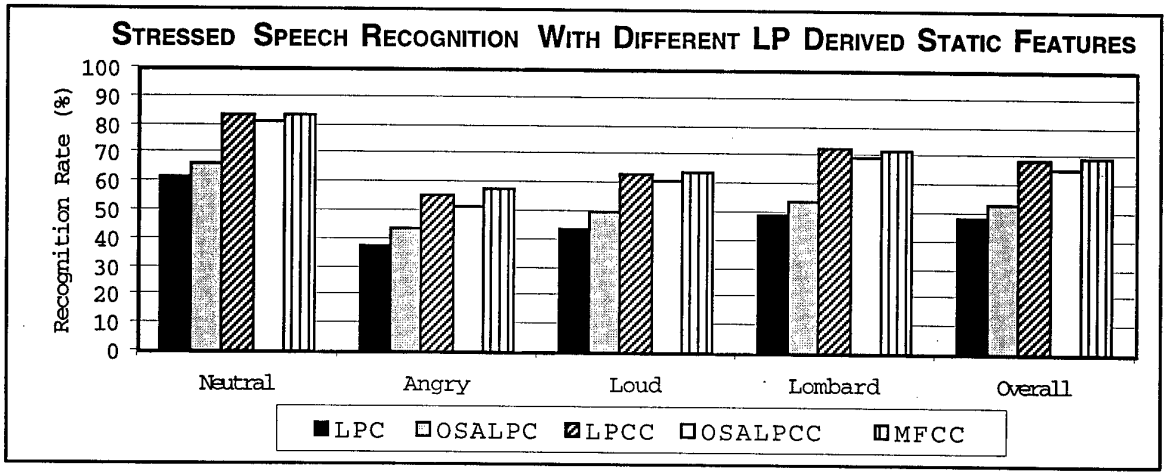


Figure 6.19: SUSAS Recognition performance of linear prediction power spectrum based static features in the presence of stress. The graph shows the recognition rates of neutral trained models tested with four speaking conditions for 5 different sets of features: LPC, OSALPC, LPCC, OSALPCC, and MFCC.

recognition.

MFCC vs. Modified Mel and Expo-Log Frequency Scales: Using a Mel-scale filter bank analysis, speech recognition across individual sub-bands were conducted across four stress conditions. The results showed that Mel-frequency cepstral coefficients (MFCC) were not always effective in recognition of stressed speech. It became evident that a new frequency scale would be needed that would emphasize mid-frequencies while de-emphasizing lower and higher frequencies. To achieve this, two new frequency partitions were proposed: one referred to as the modified mel-scale (M-MFCC), and the second, a combination of an exponential and a logarithmic function, referred to as the ExpoLog scale. Both frequency scales along with the traditional mel-scale are given below:

$$\text{mel-scale} = 2595 \times \log(1 + \frac{f}{700}) \tag{6.5}$$

$$\text{Modified mel-scale} = 3070 \times \log(1 + \frac{f}{1000}) \tag{6.6}$$

$$\text{ExpoLog} = \begin{cases} 700 \times (10^{\frac{f}{3988}} - 1) & 0 \leq f \leq 2000Hz \\ 2595 \times \log(1 + \frac{f}{700}) & 2000 < f \leq 4000Hz \end{cases} \tag{6.7}$$

Here, the ExpoLog mapping filter bank are highly concentrated at mid frequencies and sparsely distributed at frequencies below 750 Hz and above 2000 Hz. Using SUSAS speech, results from an evaluation of the three frequency warping scales for cepstral parameters were obtained (MFCC, M-MFCC, ExpoLog). Each scaling method was evaluated with a total of 2160 open test tokens. When static features were employed for recognition, M-MFCC outperformed traditional MFCC by 4.5 % for angry, 1.9 % for loud, and 5.4 % for Lombard effect. The performance of ExpoLog static features also outperformed the mel-scale, for all stress styles, with an average performance improvement of 4.8 %. Note that for angry and loud speech recognition, ExpoLog exceeded MFCC by as much as 7.6 % and 7.8 %. These results clearly showed that with a slight modification in the manner in which cepstral parameters were obtained, recognition performance in stressed speech conditions was improved.

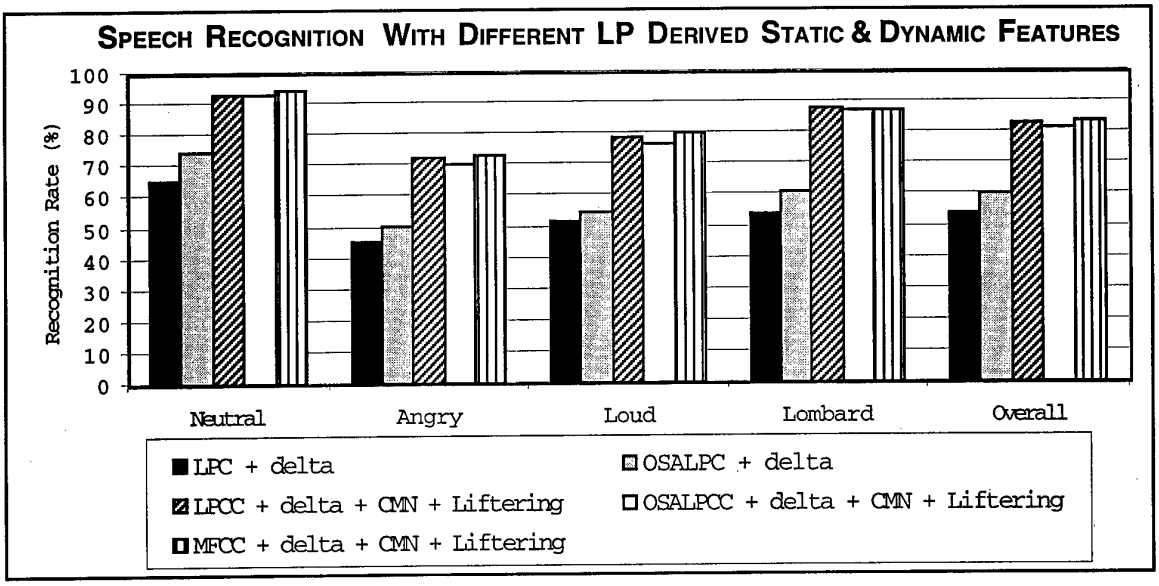


Figure 6.20: SUSAS Recognition performance of linear prediction power spectrum based static and dynamic features in the presence of stress. The graph shows the recognition rates of neutral trained models tested with four speaking conditions for 5 different sets of features: LPC, OSALPC, LPCC, OSALPCC, and MFCC.

Stressed Speech Recognition using FFT vs. Linear Prediction Power Spectrum: In a recent survey by Picone [127] of contemporary recognition systems, it was established that FFT-based spectral parameters are preferred to LP-based parameters since they are believed to be more immune to the presence of noise. Only a third of all the surveyed systems employed LP-derived parameters, the remainder used FFT based processing. For this reason, a number of systems rely on the Fourier transform-based filter bank analysis. In order to evaluate the FFT's immunity to stress, two recognition evaluations were conducted using parameters derived from FFT and LP power spectral estimation methods. An additional recognition evaluation employing actual stressed speech produced in a noisy environment was performed in order to determine which power spectral estimation method was more robust to the presence of both noise and stress. The noise in this case represented time varying mechanical and wind noise obtained from speech recorded during amusement park roller coaster rides.

The results showed that, contrary to their noise immunity, FFT-based spectral parameters were not equally robust to the presence of stress. A comparison of the performance of LP and FFT power spectrum based features is shown in Figure 6.21. For neutral training and testing, FFT based parameters performed slightly better than cepstral parameters derived from an LP spectrum. However, the LP power spectrum performed significantly better than the FFT power spectrum when neutral trained models were tested with angry, loud, and Lombard effect speech. Modified MFCC (M-MFCC) and ExpoLog based features consistently outperformed MFCC parameters using both FFT and LP based spectra, but LP derived ExpoLog produced the highest recognition rates across stressed styles using static features. Next, extending the static features to include time derivatives and feature processing, these combined parameters were shown to greatly enhance the performance of stressed speech recognition [74]. Having established the ExpoLog frequency scale as being superior to mel and modified-mel scales, the performance of ExpoLog static and dynamic features showed that the LP based features outperform FFT by an overall 3.9 %. For angry speech recognition, the difference in recognition was as high as 9.6 %.

A second evaluation was performed using actual noisy stressful speech from the SUSAS

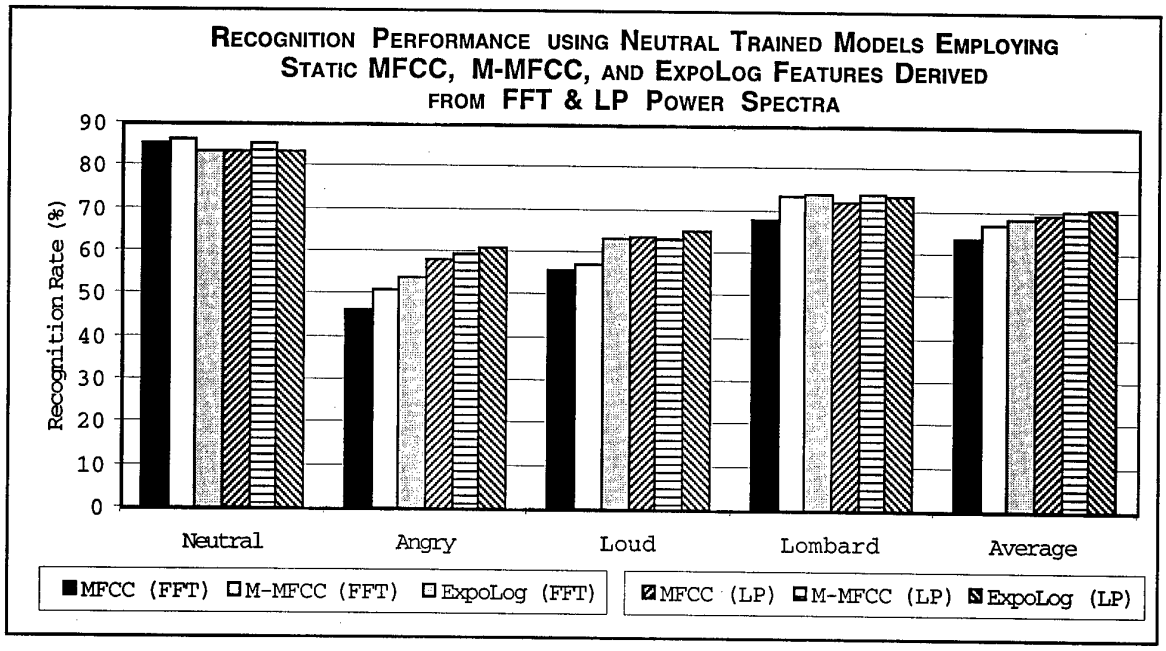


Figure 6.21: A comparison of the performance of FFT vs. LP power spectrum derived features of neutral trained models using static features. The performance of MFCC is compared to two different frequency scales.

database. This evaluation was intended to determine which power spectral estimation method was most effective when speech was subjected to a combination of noise and stress. The results, as summarized in Figure 6.22, indicate that the LP-based features outperformed the FFT-based features not only for noise-free simulated stress conditions but also for noisy actual stressed speech. It is believed that the spectral smoothing inherent in the LP model provides a more overall smooth set of parameters capable of reducing the fine fluctuations caused by excitation changes (i.e., pitch structure) that exist under stressful conditions.

Basic Parameter Processing: Preemphasis, CMN, Cepstral Liftering While it is desirable to formulate speech features which are inherently robust to the variability of speech under stress, there are a number of possible subsequent parameter processing methods which have been shown to be effective for noise and communication channel effects. It is noted that other methods, such as stress equalization feature processing (MCE-ACC [58] and others in [60]), have been shown to be effective in reducing the impact of stress. However, such stress equalization processing requires stress and/or word dependent compensation terms. The goal here is to consider only feature processing methods which do not require knowledge of either word or phoneme class sequence content, or the type of speaker stress. For these experiments, the following three parameter processing methods were considered: preemphasis, liftering, and cepstral mean normalization, which have been widely used for improved speech recognition and speaker identification. Here, their contribution to stressed speech recognition was evaluated.

Fixed and Slowly-Varying Preemphasis Previous analysis studies on stressed speech have shown that the spectral structure and overall average spectral slope varies for different speaking conditions [54, 60, 160, 136, 143, 147]. Since the average spectral slope of the input speech is different for various stressed speaking styles, then in order to flatten the spectral tilt, it is necessary to vary the filter parameters according to the input speech. Therefore, it is proposed to use an adaptive preemphasizer where only the spectral slope of voiced speech is adaptively

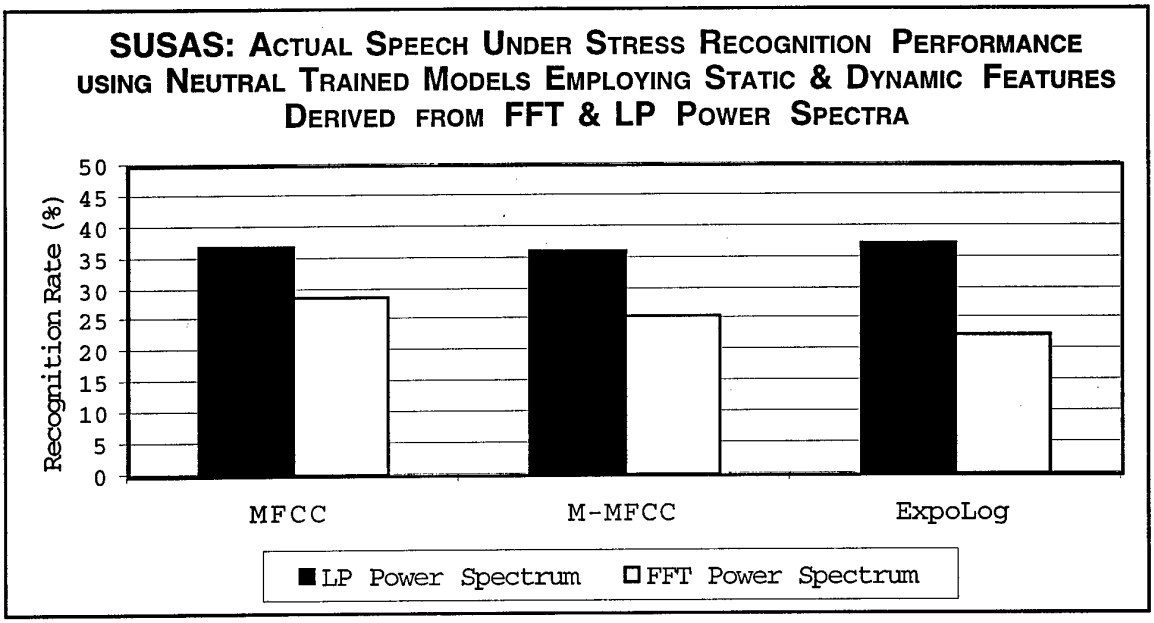


Figure 6.22: Recognition performance of neutral trained models employing static and dynamic MFCC, M-MFCC, and ExpoLog features derived from FFT and LP power spectra in noisy actual stressed conditions.

flattened while the unvoiced speech sections are not preemphasized. The adaptive preemphasizer chosen was a slowly-varying first order filter [130]. The variable filter coefficient was represented as a ratio of the first to the zeroth order lag autocorrelation parameters. The filter was applied to utterances both during training and testing. An evaluation using SUSAS data showed that the slowly-varying preemphasis filter improved recognition of angry speech by 2.4 % and that of Lombard effect speech by 1.5 %.

Fixed and Variable Cepstral Mean Normalization In a study of channel compensation techniques for speaker identification, simple cepstral mean removal was the best channel compensation method when compared to RASTA processing and quadratic trend removal [132]. Cepstral mean normalization is a simple yet effective method which assumes no knowledge of the environment and is employed for reducing long term differences in channel characteristics. Since the presence of stress impacts voiced and unvoiced speech phonemes differently [54, 71], it is proposed to compute a separate cepstral mean for voiced and unvoiced sections instead of computing a single mean across the entire utterance. A series of the SUSAS recognition evaluations were performed using MFCC static parameters, with various configurations of delta parameters, fixed or variable preemphasis, and fixed or variable cepstral mean normalization (CMN). The results, shown in Figure 6.23, indicate that variable cepstral mean normalization performed better than traditional CMN when no delta parameters were employed. The recognition of angry, loud, and Lombard effect speech was improved respectively by 4.1 %, 3.5 % and 1.5 %. Variable CMN was most effective with static features and fixed preemphasis.

Cepstral Liftering The last feature processing method considered was cepstral liftering. Cepstral liftering is a weighting technique applied to cepstral coefficients in order to reduce the spectral slope or the undesirable broadband noise components of the spectrum, which affect low order cepstral parameters, while retaining the essential characteristics of the formant structure. The low-order cepstral coefficients are believed to be primarily sensitive to overall spectral slope,

speaker characteristics, or vocal efforts. The higher order cepstral coefficients represent fine spectral structure. Cepstral liftering was also evaluated using MFCC parameters for SUSAS speech (see Figure 6.23). The evaluations showed that when cepstral liftering was employed with time derivative parameters and cepstral mean normalization, it had no effect on the recognition performance of neutral trained models tested with neutral, angry, loud, and Lombard conditions. Variable preemphasis slightly improved recognition over fixed preemphasis. Finally, variable CMN improved recognition over fixed CMN by 3 % when static parameters were employed for recognition. Their performance was not equally effective when time derivative parameters were included. The final recommendation from the RPSL study on robust features was that for effective speech recognition performance in both neutral and stressed conditions, speech recognizers should (i) employ features derived from an LP as opposed to an FFT based power spectrum, and (ii) use a modified frequency partition such as M-MFCC or ExpoLog if possible. In addition, it was reported that variable preemphasis and variable CMN both improved stressed speech recognition performance, but that their impact was reduced if time derivative parameters were also included.

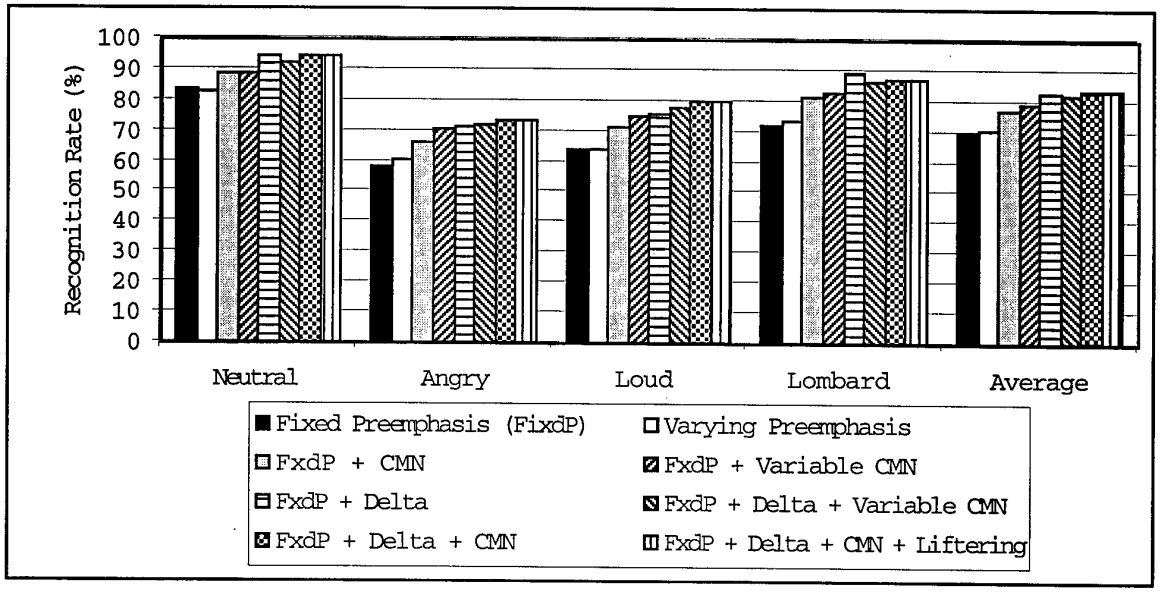


Figure 6.23: Effect of preemphasis (fixed and variable), cepstral mean normalization (fixed and variable), time-derivative (Δ coefficients), and cepstral liftering on the recognition performance of LP based MFCCs. Note, "FxdP" refers to "Fixed Preemphasis", and "Delta" refers to time-derivative delta coefficients.

6.2.4 Automatic Speech Recognition Conclusions

In general, we see a degradation in the performance of recognition systems when the user is under stress, regardless of the system under test. In the case of the DLP database the relationship to increasing speaker stress is clear as the increasing frequency of the plate reading task corresponds to an increase in user stress, although if we look at the DERA results for individuals (see Figure 6.3), we see that this is not always the case for a given individual. It was also evident that the level of task stress experienced by the speaker was mild, given the potential range of stress an operator may face in military environments.

The range and levels of stress displayed within the SUSAS database was more extensive. The simulated stress domains of SUSAS have speakers producing speech under different emotions or computer response tasks, and therefore the stress levels were not as significant as the actual stress

domains (roller-coaster rides, helicopter pilot speech, etc.). In tests using the SUSAS database the correlation is not so clear. Results of tests on the different simulated and actual stressed conditions show a clear degradation from neutral speech and we can see that the commercial off-the-shelf recognition systems tested do not perform as well when speech is under stress. For example, all systems tested demonstrated the lowerest recognition results when the simulated angry speech was used.

As we would expect, large vocabulary speaker-independent task-independent (COTS) systems did not do as well as systems with vocabularies limited to that used in the datasets or those trained on the subsets of the database. Previous methods which use multi-style training can improve performance, but this has been shown to work only for speaker-dependent recognizers. The results for the DERA speaker-independent task-independent 35 word vocabulary system performed better than several systems trained on the database in some conditions (see Figure 6.15), indicating the progress being made in these systems.

The compensation techniques described in the last section have also shown improvements in the recognition rate for the system tested and reduced the standard deviation of the rates, (i.e. the results were more consistent across different stress speaking conditions). Those methods which directly compensate for stress in the speech recognition features, typically at the phone-class level, are more effective than general compensation such as whole word cepstral mean normalization. However, it is also known that speakers will often differ in how they display stress in their speech under these conditions, thereby suggesting that speaker-dependent approaches may be the only viable solution. It is noted that some techniques reduced the number of insertion errors.

From those studies which considered overall feature processing, the recommendation is to use (i) an LP based spectral feature set instead of one based on an FFT, (ii) the use of a modified frequency partition such as a modified mel-frequency scale coefficients or ExpoLog, and (iii) the use of second order parameters.

6.3 Speaker Recognition and Verification

There has been great interest in military voice communications, command, and control for reliable speaker recognition. These involve both identification (input speaker is 1 of N known speakers) and verification (input speaker claims to be 1 speaker, but could come from an unrestricted population). A number of speaker recognition studies include those by Reynolds, et al. [132, 133, 134], and others [135, 48, 32]. The issue of reliable speaker identification under stressful speaking conditions however has received little attention. This is primarily due to the lack of a large, well organized database for such studies. Databases such as the YOHO[2] or NIST Switchboard[3] databases for example, are produced under calm neutral speaking conditions. In this section, we discuss speaker recognition evaluations conducted by RMA/SIC on the SUSC-0 database, and by RSPL on the SUSAS database.

6.3.1 Evaluations Conducted Using SUSC-0

Tests were conducted at the Speech Lab of the Royal Military Academy (Belgium) (RMA/SIC) using the speech material in the SUSC-0 database to determine the effects of stressed speech on the results of a speaker recognition system.

The number of different speakers needed to measure the performance of speaker recognition systems has to be reasonably high, so, the SUSC-0 database was chosen. Nine different controllers appear in the database, but as the duration of the speech uttered by speaker VL was too

[2]YOHO is a 138 speaker database where speakers produced 3-pair digit combination lock sequences, and has been used extensively in algorithm development for speaker recognition.

[3]The NIST Switchboard database consists of over 500 speakers talking extemporaneously over telephone channels and has been used for competitive evaluations since 1996.

short, all measurements were made with the eight other speakers. The sentences uttered by the controllers were different from one speaker to another; moreover, the "non-stressed sentences" were not in the same language as the "stressed sentences". As a result, it was only possible to use speaker recognition methods that are text- and language-independent.

Speaker Recognition System The speaker recognition system used in these experiments was the Vector Auto-Regressive (VAR) method. In this method, each speaker is characterized by two prediction matrices which are used in a second order vector-equation to predict the sequence of cepstral vectors from a speech file.

During the training phase, the matrices were determined so as to minimize the total quadratic prediction error of the considered vector sequence obtained from 20 seconds of speech. For an identification test, the prediction matrices from each speaker were used to calculate the total quadratic prediction error related to the test sequence. The identified speaker was the one with the lowest total quadratic error. For a verification test, only the prediction matrices of the claimed speaker were used in order to calculate the total quadratic error of the given sequence. This error was then compared with a threshold to make a decision about the speaker.

The speech sequence required for this method has to be of adequate length. The training database contained four sentences for each speaker. These sentences were obtained by concatenation of speech material from each speaker in order to obtain the following durations. For the first three sentences, the duration was about 20 seconds and for the fourth sentence, the duration was about 40 seconds. This last sentence was obtained by the concatenation of the first two sentences and permitted the study of the effect of the duration of the speech sequence on the results. With these four sentences, it was possible to establish four models for each speaker.

The test database contained 10 sentences per speaker. Each of them was also obtained by concatenation of the available speech material in order to have a duration of about 7 seconds.

Due to the existence of two recording conditions, it was possible to conduct three test cases:

1. training with neutral speech, test with neutral speech
2. training with neutral speech, test with stressed speech
3. training with stressed speech, test with stressed speech

Results of the identification tests For each test case, four identification tests of the 8 speakers were completed. Each test corresponds to one of the four speaker's models.

Figure 6.24 summarizes the results of the experiments. The results show that there was a large increase in the error rate when the tested speech condition did not correspond with the training condition, however, when the speech was stressed, even if the training condition corresponds to the test condition, there was an increase in the identification error rate.

Results of the verification tests A determination of the false acceptance rates and of the false rejection rates was made for the eight speakers, using the four models for each speaker. Figure 6.25 and 6.26 are graphs for the first speaker. Figure 6.25 gives the false acceptance rate (rejection-curve) and the false rejection (verification-curve) as a function of the decision threshold obtained with the speaker's model 1 for the three test cases. Figure 6.26 is the corresponding ROC-curve.

The first graph is very interesting because it shows that the three curves of the false acceptance rate are independent of the test case. Whether the speech was stressed or not does not change this particular error rate. The value obtained relates more to the speaker recognition method used than to the speech model. This was not true for the false rejection rate, here the results for test case 2 were not as good as those of the other test cases.

If the same results are represented with ROC-curves (Figure 6.25), we can see that the test case 2 results were bad in comparison with those of test case 1 and test case 3. However it is

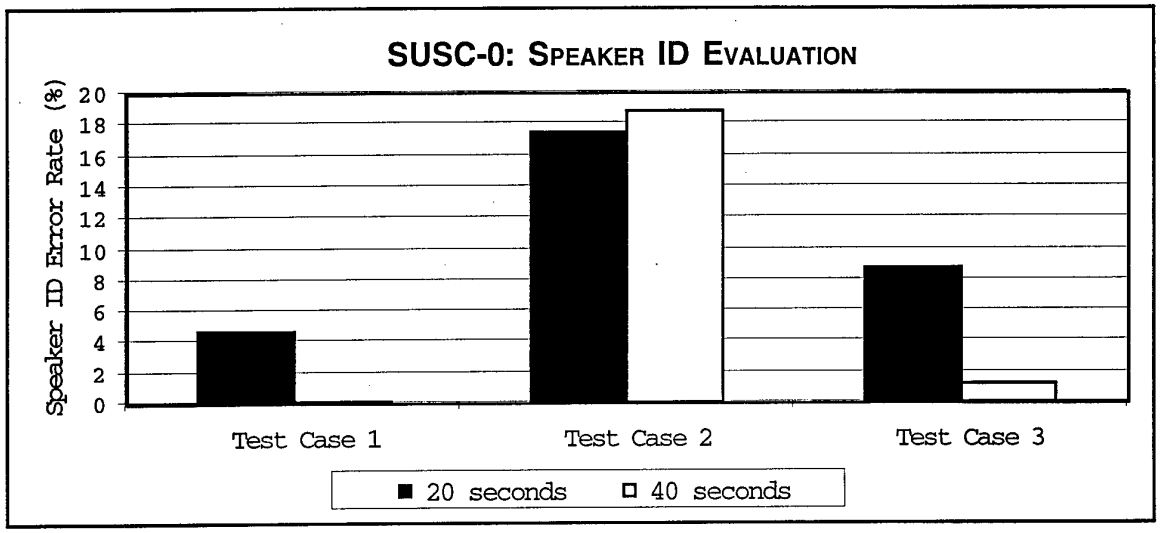


Figure 6.24: Results of the Speaker Identification Tests (Note: error rate in test case 1 for 40 seconds test sample is 0 %.)

not possible with this type of curve to emphasize the independence of the false acceptance rate (rejection-error) with respect to the test case.

Conclusion Due to the concatenation of the speech material to obtain different sentences with a sufficient duration, it was not possible to relate the stress level of the speakers to the sentences used, therefore, the results of this study are rather rough. However, they permit us to conclude that stress does have an influence on speaker recognition results.

To identify stressed speakers, it is insufficient to have a non-stressed model of the registered persons. It is thus interesting to build a "stressed model" of each person, leading in this way to a good identification and verification of the stressed speaker. The remaining problem in a real case would be to determine whether the tested speaker is stressed or not.

6.3.2 Evaluations Conducted Using SUSAS

Tests were conducted by RSPL, using speech material from the SUSAS database to determine the effects of speech under stress on a speaker recognition system.

Speaker Identification System and Test Set-up: A standard Gaussian mixture model (GMM) speaker identification system was implemented [132]. The GMM system employed 32 Gaussian mixture weights, with a parameter set consisting of 19 Mel-frequency cepstral coefficients (MFCC). Here, a preemphasis was first performed with a coefficient of 0.97, and the parameters were found using 20 filterbank channels. We also note that $c[0]$ was excluded. The parameters were obtained using an analysis window width of 20ms, with a skip rate of 10 ms. The training data consisted of 35 isolated words per speaker, per stress condition. Nine SUSAS speakers across 7 stress conditions were used to evaluate the effects of speaker stress on speaker identification accuracy.

Experiment 1: Here, we want to see the effects of mismatch in speaker stress style on speaker identification. As such, neutral trained models are used to identify speakers with stressed speech. We note that this was an open test.

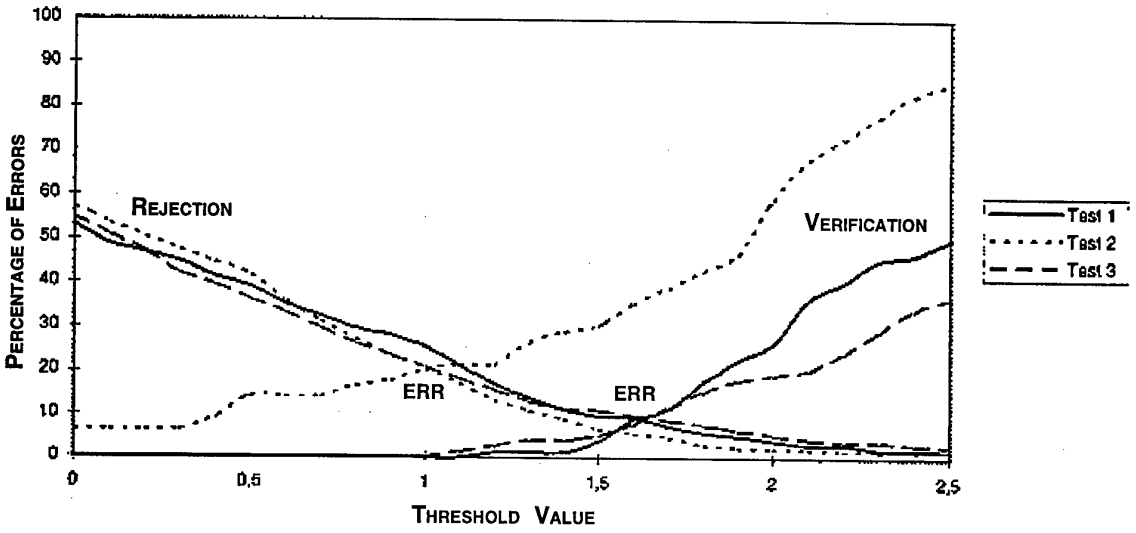


Figure 6.25: Verification and Rejection errors.

Results from Experiment 1: The results summarized in Figure 6.27 illustrate how speaker ID performance can quickly degrade as a result of variations in speaker stress.

Experiment 2: This experiment is intended to determine if cepstral-mean normalization will improve or degrade speaker identification in mismatched training/testing conditions. Cepstral-mean normalization has been used to remove long-term spectral structure, such as microphone or channel characteristics, for speaker recognition. For this experiment, the speaker models were retrained using mean normalized features and mean normalized testing data. This was also an open test experiment.

Results from Experiment 2: The results summarized in Figure 6.28 show speaker ID performance with cepstral-mean normalization applied to training and test data. Here we see that cepstral-mean normalization actually leads to poorer performance over models trained and tested without cepstral-mean normalization.

6.3.3 Discussion

Although the databases available do not contain enough material to do comprehensive speaker identification or verification evaluations under stressful conditions, several preliminary experiments have been carried out on two of the databases described in this report. The results of the experiments indicate that speech spoken under stressful conditions degrade the performance of speaker recognition systems. The performance observed when testing a system which was trained with neutral speech can be less than half those of a system trained with material collected under the same stressed conditions.

6.4 Stressed Speech Synthesis and Coding

A limited number of studies have integrated stressed speech variations in speech synthesis systems to improve the naturalness of synthetic speech [2, 22, 118, 139]. Previous approaches directed at integrating emotion in text-to-speech synthesis systems have concentrated on formulating a set of fixed rules to represent each emotion. However, analysis studies on emotion and stress suggest that using a fixed set of rules would ultimately represent merely a single caricature of speech variations under a certain emotional condition rather than representing

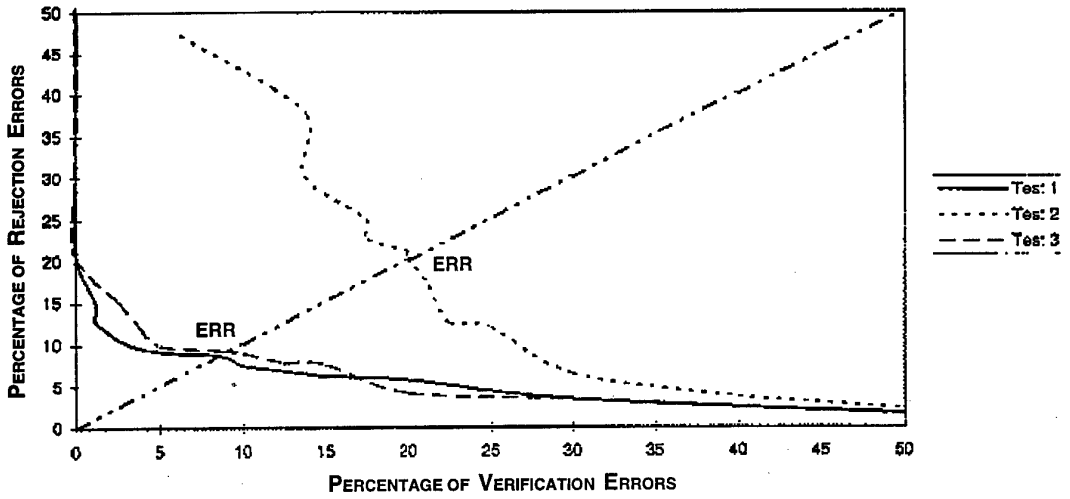


Figure 6.26: ROC Curve.

the range of variations for continuous speech that may exist under stress. A stressed speech parameter modeling and perturbation scheme based on a code-excited linear prediction (CELP) vocoder was previously employed for speaking style modification of neutral speech [16]. While the speech parameter perturbation within a CELP framework was effective and successful based on a formal listener assessment, the approach was text-dependent and restricted to the vocoder's framework.

Some researchers have suggested that "no commercial TTS system incorporates prosodic variation resulting from emotion and related factors" [119]. In fact, the voice quality of current TTS systems is still easily distinguishable from natural voices and introducing pragmatic effects has always been of lower importance than intelligibility. Thus, the study of vocal emotions has been conducted mostly in academic environments by physiologists on one side and speech researchers on the other.

Progress in this area has been slow, partly due to the complexity of identifying and categorizing the emotion factors in human natural speech, and implementing these factors within synthetic speech. It is known that emotion causes changes in voice quality, pitch contour and speech rate.

A few prototype TTS systems include some sort of vocal emotion synthesis capability, by manipulation of the above mentioned parameters. Examples are: the HAMLET system by Murray [110], the Affect Editor system by Cahn [22], and the SPRUCE system by Tatham [155].

The European project VAESS (Voices, Attitudes and Emotions in Speech Synthesis—Tide Program) also approached this subject. In this context, a prototype system for Spanish has been enhanced with the capability of simulating three emotions (anger, happiness and sadness).

Emotions and other stressors have been studied in a totally different context by Bou-Ghazale and Hansen [16]. The stress perturbation algorithm has been formulated based on a CELP coding structure, for isolated words under neutral, loud, angry and Lombard effect speaking conditions. This study showed that the perturbations of neutral speech that better conveyed the emotional state of the speaker were a combination of pitch, gain and formant location modifications. A later formulation, generalized the approach using hidden Markov models to represent the stress deviation from neutral (Bou-Ghazale and Hansen [19]). This method demonstrated that it was possible to model the changes which occur in stressed speech production from a given training speaker set, and use this knowledge and model representation to perturb unseen

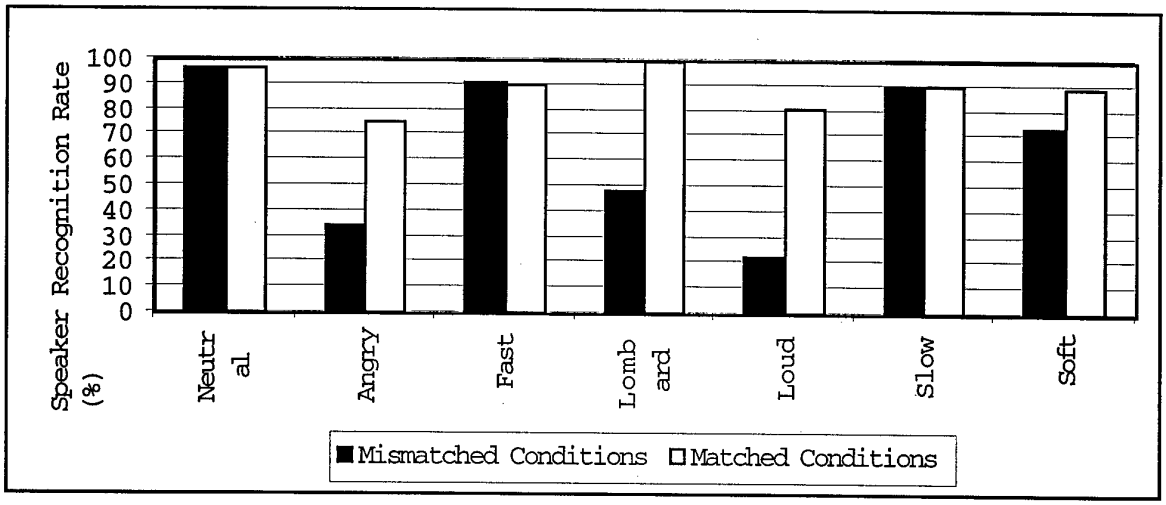


Figure 6.27: Speaker recognition results (in %) from a GMM system trained using neutral speech data and tested with speech under stress (mismatched condition), and GMM system trained and open tested with speech under stress (matched condition).

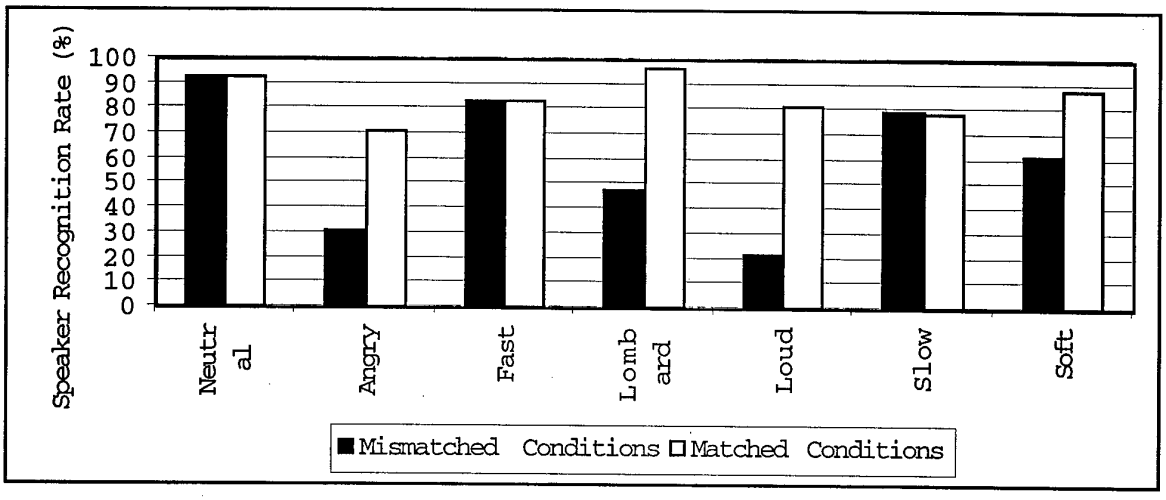


Figure 6.28: Speaker recognition results (in %) from a GMM system with cepstral-mean normalization (CMN), trained using neutral speech data and tested with speech under stress (mismatched condition), and GMM system trained and open tested with speech under stress (matched condition).

neutral input speakers so their output reflected certain stress speaking styles.

Despite the slow general progress in this area, it is worth mentioning that prosodic content in synthetic speech is seen as increasingly important, and there is presently renewed interest in the derivation of models of emotion.

6.5 Conclusions

The results in of the experiments in this chapter have shown significant degradation in performance of commercial off-the-shelf speech and speaker recognition systems when the talker is under stress. Techniques to compensate for the stress have only been marginally effective. Continued research into the techniques to reduce the impact of stress is required, especially for applications in military environments.

6.6 Selected References of Interest:

Here, we summarize several references which have considered areas of recognition, speaker identification, and/or synthesis of speech under stress. The reference section at the end of this report contains all references cited in this chapter.

Recognition

1. D.A. Cairns, J.H.L. Hansen, "ICARUS: An Mwave Based Real-time Speech Recognition System in Noise and Lombard Effect," *ICSLP-92, Inter. Conf. Spoken Lang. Proc.,* pp. 703–706, Alberta, Canada, October 1992.

2. Y. Chen, "Cepstral Domain Talker Stress Compensation for Robust Speech Recognition," *IEEE Trans. on Acoust., Speech, Sig. Proc.,* pp. 433–439, April 1988.

3. J.H.L. Hansen, "Morphological Constrained Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect," *IEEE Trans. Speech & Audio,* SPECIAL ISSUE: Robust Speech Recognition, **2**(4):598–614, 1994.

4. J.H.L. Hansen, S. Bou-Ghazale, "Duration and Spectral Based Stress Token Generation for Keyword Recognition Using Hidden Markov Models," *IEEE Trans. Speech and Audio Proc.,* vol. 3(5), pp. 415–421, Sept. 1995.

5. J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition," *Speech Communications, Special Issue on Speech Under Stress,* vol. 20, pp. 151–173, Nov. 1996.

6. B.A. Hanson, T. Applebaum, "Robust Speaker-Independent Word Recognition Using Instantaneous, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech," *IEEE 1990 ICASSP,* pp. 857–60, Apr. 1990.

7. R.P. Lippmann, E.A. Martin and D.B. Paul, "Multi-Style Training for Robust Isolated-Word Speech Recognition," *IEEE ICASSP-87,* pp. 705–708, 1987.

8. B.A. Mellor, R. Graham, "The Effect on Speech Recogniser Performance of Cognitive Stress Induced by a Time Constrained Task," *DRA Memo #4749,* Unreleased, Sept. 1993.

9. D.B. Paul, "A Speaker-Stress Resistant HMM Isolated Word Recognizer," *IEEE 1987 ICASSP,* pp. 713–716, 1987.

10. P.K. Rajasekaran, G.R. Doddington, J.W. Picone, "Recognition of speech under stress and in noise," *IEEE 1986 ICASSP,* pp. 733–736, 1986.

11. B.J. Stanton, L.H. Jamieson, G.D. Allen, "Robust Recognition of Loud and Lombard Speech in the Fighter Cockpit Environment," *IEEE 1989 ICASSP,* pp.675–8.

Synthesis:

1. J.L. Arnott, A.F. Newell, I.R. Murray, E. Abadjieva, "Development of rule systems for the simulation of mood and emotion in speech synthesis," *Final Project Report to the SERC,* Dundee University, Microcentre, July 1993, pp. 1–33.

2. S. Bou-Ghazale, J.H.L. Hansen, "Stressed Speech Synthesis Based on a Modified CELP Vocoder Framework," *Speech Communications: Special Issue on Speech Under Stress,* vol. 20, pp. 93–110, Nov. 1996.

3. S.E. Bou-Ghazale, J.H.L. Hansen, "Stress Perturbation of Neutral Speech for Synthesis based on Hidden Markov Models," *IEEE Trans. on Speech & Audio Processing,* vol. 6, no. 3, pp. 201–216, May 1998.

4. J. Cahn, "The generation of affect in synthesised speech", *Journal of the American Voice I/O Society,* Vol. 8, pp. 1–19, 1990.

5. K.E. Cummings, M.A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," *J. Acoust. Soc. Am.,* **98**(1) 88–98, 1995.

6. I.R. Murray, J.L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. of America,* 93(2), pp. 1097–1108, Feb. 1993.

7. I.R. Murray, J.L. Arnott, E.A. Rohwer, "Emotional stress in synthetic speech: Progress and future directions," *Speech Communication,* Vol. 20, Nos. 1–2, Nov. 1996.

Chapter 7

Conclusions & Recommendations

The field of military speech technology requires the integrated use of speech systems for communications, command, and control. In addition, for multi-national environments, it is necessary for a wide range of protocols from participating countries to be integrated together for safe and effective operations. Speech technology in military environments offers the promise of more direct and effective communciations, verification of personnel, and allowing operators to have access to better information. The problems of battlefield stress conditions, however, raise a serious obstacle for the transition of commercial off-the-shelf speech technology for speech recognition, speaker verification, synthesis and coding. Studies conducted by participating NATO laboratories and discussed here suggest that many COTS speech systems which were designed for quiet or low-noise office environments, cannot be effectively used in real-world, high task stress, emotional induced, high background noise, and operator fatigued situations. The benefit of voice technology in such environments is clear; however the advances in basic research needed to address these environments has not kept up with demand for effective solutions. It is suggested that this report will serve as useful vehicle to focus the speech community (both industry, academia, government and military labs) on important issues of speech variability under stress. Databases obtained or collected during this study have been distributed to all participating NATO countries, and several are available in CD-ROM format for those interested[1]. Below, we summarize the main findings and recommendations.

1. Military operations are often conducted under conditions of stress induced by high workload, sleep deprivation, fear and emotion, confusion due to conflicting information, psychological tension, pain, and other typical conditions encountered in the modern battlefield context. These conditions are known to affect the physical and cognitive abilities of human speech characteristics.

2. It is suggested that operator based stress factors are likely to be detrimental to the effectiveness of communication in general, as well as to the performance of communication equipment and weapon systems equipped with vocal interfaces (e.g., advanced cockpits, C^3—command, control, and communication systems, information warfare).

3. Commercial off the shelf speech recognition systems are not yet able to address the wide speaker variability associated with speech produced under stress.

4. Progress in the field of military based speech technology, including advances in speech based system design has been restricted due to the lack and availability of databases of speech under stress. In particular, the type of stress which an operator may experience in the modern battlefield context is not easily simulated, and therefore it is difficult to systematically collect speech data for use in research and speech system training.

[1]The SUSAS Stressed Speech Database from RSPL is available from the Linguistics Data Consortium at the following web location: `http://morph.ldc.upenn.edu/Catalog/LDC99S78.html`

5. It is certain that in the future it will be more necessary to improve the coordination of multi-national military forces. The need therefore exists for planned simulations requiring co-ordinated emergency or military personnel using a wide range of speech technology. Such battlefield settings will have to address factors an operator would actually experience such as high workload, sleep deprivation, fear and emotion, confusion, psychological tension, pain, etc.

6. The success of this three year effort by RSG10 (now IST/TG-01) has underlined the necessity to further invest coordinated international effort to support NATO interests in understanding speech production and perception and our ability to implement speech systems which are robust to the specific realities of everyday military speech.

7. In order to share the most recent advances in this field, then NATO IST/TG-01 established their Speech Under Stress web-page, which is found at the following Web location: `http://cslu.colorado.edu/rspl/stress.html` Information found here include an overview of our activities, namely collected and available speech databases (with audio demonstrations), international research groups, and an extensive set of references. A copy of this report is also available at that web page.

Bibliography

[1] *Speech Communication*: Special Issue on Speech under "Stress", Vol. 20, Nos. 1–2, Nov. 1996.

[2] E. Aadjieva, I.R. Murray, J.L. Arnott, "An enhanced development system for emotional speech synthesis for use in vocal prostheses," *Proc. of ECART 2, 2nd European Conference on the Advancement of Rehabilitation Technology,* Stockholm, Sweden, 26–28 May 1993. paper 1.2, pp. 4–6.

[3] E. Abadjieva, I.R. Murray, J.L. Arnott, "Applying analysis of human emotional speech to enhance synthetic speech"," *Proc. Eurospeech'93, 3rd European Conf. on Speech Communication and Technology,* Berlin, Germany, 21–23 September 1993, pp. 909–912.

[4] T.R. Anderson, T.J. Moore and R.L. McKinley, "Issues in the development and use of speech recognition data base for military cockpit environments," *Proc. of Speech Tech. '85, Media Dimensions,* 172–176, 1985.

[5] J.L. Arnott, N. Alm, I.R. Murray, "Enhancing a communication prosthesis with vocal emotion effects," *Proceedings of the ESCA Tutorial and Research Workshop, Speech and Language Technology for Disabled Persons,* Stockholm, Sweden, 31 May-2 June 1993 pp. 165–168.

[6] J.L. Arnott, A.F. Newell, I.R. Murray, E. Abadjieva, "Development of rule systems for the simulation of mood and emotion in speech synthesis," *Final Project Report to the SERC,* Dundee University, Microcentre, July 1993, pp. 1–33.

[7] L.M. Arslan, "Foreign Accent Classification," Ph.D. Thesis, Robust Speech Processing Laboratory, Duke University, Department of Electrical and Computer Engineering, July, 1996.

[8] L.M. Arslan and J.H.L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *J. Acoust. Soc. America,* vol. 102, no. 1, pp. 28–40, July 1997.

[9] P. Benson, "Analysis of the Acoustic Correlates of Stress from an Operational Aviation Emergency," Proc. ESCA-NATO Tutorial and Research Workshop on Speech Under Stress, Lisbon, Portugal, pp. 61–64, 1995.

[10] Z.S. Bond, T.J. Moore, and B. Gable, "Some phonetic characteristics of speech produced in noise," *Journal of Acoustical Society of America,* S49 (A), 1986.

[11] Z.S. Bond, T.J. Moore, and T.R. Anderson, "The effects of high sustained acceleration on the acoustic-phonetic structure of speech: A preliminary investigation," *Journal of the American Voice I/O Society,* 4, 1–19, 1987.

[12] Z.S. Bond, T.J Moore, "Effect of Whole Body Vibration on Acoustic Measures of Speech," *Space and Environmental Medicine,* pp. 989–993, Nov. 1990

[13] Z.S. Bond, T.J. Moore, "A note on Loud and Lombard Speech," *1990 ICSLP,* pp. 969–972.

[14] S.E. Bou-Ghazale, "Duration and Spectral based Stress Token Generation for Keyword Recognition using Hidden Markov Models," M.S. Thesis, Robust Speech Processing Laboratory, Duke Univ., Dept. of Electrical Engineering, June 1993.

[15] S.E. Bou-Ghazale, J.H.L. Hansen, "Duration and Spectral Based Stress Token Generation For HMM Speech Recognition under Stress," *IEEE 1994 ICASSP,* pp. 413–416.

[16] S.E. Bou-Ghazale, J.H.L. Hansen, "Stressed Speech Synthesis Based on a Modified CELP Vocoder Framework," *Speech Communications: Special Issue on Speech Under Stress,* vol. 20, pp. 93–110, Nov. 1996.

[17] S.E. Bou-Ghazale, "Analysis, Modeling, and Perturbation of Speech Under Stress with Applications to Synthesis and Recognition," Ph.D. Thesis, Robust Speech Processing Laboratory, Duke Univ., Dept. of Electrical Engineering, November 1996.

[18] S.E. Bou-Ghazale, J.H.L. Hansen, "Speech Feature Modeling Approaches for Robust Speech Recognition Under Stress," submitted to *IEEE Trans. on Speech & Audio Proc.*, 24 pgs, Oct. 1997. Revised Feb. 1999.

[19] S.E. Bou-Ghazale, J.H.L. Hansen, "Stress Perturbation of Neutral Speech for Synthesis based on Hidden Markov Models," *IEEE Trans. on Speech & Audio Processing,* vol. 6, no. 3, pp. 201–216, May 1998.

[20] O. Bria, "Improved Automatic Speech Recognition Under Lombard Effect," M.S. Thesis, Robust Speech Processing Laboratory, Duke University, Dept. of Electrical Engineering, April 1991.

[21] S.R. Browning, et al, "Texts of Material Recorded in the SI89 Speech Corpus," *SP4 Research Note #142,* Feb. 1991.

[22] J. Cahn, ""The generation of affect in synthesised speech", *Journal of the American Voice I/O Society,* Vol. 8, pp. 1–19, 1990.

[23] D.A. Cairns, "Real-time Speech Recognition under Lombard Effect and in Noise," M.S. Thesis, Robust Speech Processing Laboratory, Duke University, Dept. of Electrical Engineering, April 1991.

[24] D.A. Cairns, J.H.L. Hansen, "ICARUS: An Mwave Based Real-time Speech Recognition System in Noise and Lombard Effect," *ICSLP-92, Inter. Conf. Spoken Lang. Proc.,* pp. 703–706, Alberta, Canada, October 1992.

[25] D.A. Cairns, J.H.L. Hansen, "Nonlinear Analysis and Detection of Speech Under Stressed Conditions," *Journal of the Acoustical Society of America,* vol.96, no.6, pp. 3392–3400, Dec. 1994.

[26] D.A. Cairns, J.H.L. Hansen, "Nonlinear Speech Analysis using the Teager Energy Operator with Application to Speech Classification under Stress," *ICSLP-94: Inter. Conf. on Spoken Lang. Proc.,* vol. II, vol. 3, pp. 1035–1038, Yokohama, Japan, Sept. 1994.

[27] M. Carey, E. Parris, and J. Bridle, "Methods and Apparatus for Verifying the Originator of a Sequence of Operations," *United States Patent,* Patent No. 5, pp. 526–465, June 11, 1996.

[28] R. Carlson, B. Granstrom, L. Nord, "Experiments with emotive speech - Acted utterances and synthesized replicas," *ICSLP-92: Inter. Conf. on Spoken Lang., Proc.,* vol. 1, pp. 671–674, Banff, Canada, 1992.

[29] V.L. Cestaro, "A Comparison between Decision Accuracy Rates Obtained Using the Polygraph Instrument and the Computer Voice Stress Analyzer (CVSA) in the Absence of Jeopardy", Tech. Report, DoD Polygraph Inst., Aug. 1995.

[30] Y. Chen, "Cepstral Domain Talker Stress Compensation for Robust Speech Recognition," *IEEE Trans. on ASSP,* pp.433–439, April 1988.

[31] G.J. Clary, J.H.L. Hansen, "A Novel Speech Recognizer for Keyword Spotting," *ICSLP-92: Inter. Conf. on Spoken Lang. Proc.,* pp.13–16, Oct. 1992.

[32] J.M. Colombi, D.W. Ruck, S.K. Rogers, M. Oxley, "Cohort selection and word grammar effects for speaker recognition," *IEEE 1996 ICASSP,* pp. 85–88, May 1996.

[33] F.S. Costanzo, N.N. Markel, P.R. Costanzo, "Voice quality profile and perceived emotion," *Journal of Counseling Psychology,* vol. 16:3, pp. 267–270, 1969.

[34] R. de Crdoba, X. Mendez-Pidal, J. Macas-Guarasa, A. Gallardo-Antoln and J.M. Pardo, "Development and Improvement of a Real-Time ASR System for Isolated Digits in Spanish over the Telephone Line", *EUROSPEECH-95: 4th European Conf. on Speech Communication and Technology,* vol. 2, pp. 1537–1540, Madrid Spain, Sept. 1995.

[35] R. Cruise, D. Denison and P. K. Rajasekaran, "Speech recognition in the helicopter vibration environment," *(Unpublished paper presented at the 1986 Human Factors Society Meeting, Dayton, Ohio,* 1986.

[36] K.E. Cummings, M.A. Clements, and J.H.L. Hansen, "Estimation and Comparison of the Glottal Source Waveform Across Stress Styles Using Glottal Inverse Filtering," *Proc. of the IEEE Southeastcon,* pp. 776–781, Columbia, South Carolina, April 1989.

[37] K.E. Cummings, M.A. Clements, "Analysis of Glottal Waveforms Across Stress Styles," *IEEE ICASSP-90: Inter. Conf. on Acoust., Speech, Sig. Proc.,* pp. 369–372, 1990

[38] K.E. Cummings, M.A. Clements, "Improvements to and applications of analysis of stressed speech using glottal waveforms," *IEEE ICASSP-92,* **II**:25–28, 1992.

[39] K.E. Cummings, M.A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," *J. Acoust. Soc. Am.,* **98**(1) 88–98, 1995.

[40] J.K. Darby, *Speech Evaluation in Psychiatry,* Grune & Stratton, New York, New York, 1981.

[41] B.A. Dautrich, L.R. Rabiner, T.B. Martin, "On the effects of varying filter bank parameters on isolated word recognition," *Trans. on Acoustics, Speech, and Signal Processing,* vol. ASSP-31, no. 4, pp. 793–806, August 1983.

[42] S.B. Davis, P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustics, Speech, and Signal Proc.,* vol. ASSP–28, pp. 357–366, 1980.

[43] T. Dennison, "The effect of simulated helicopter vibration on the accuracy of a voice recognition system," *In Proceedings of the American Helicopter Society Annual Forum and Technical Display,* 1, 133–136, 1985.

[44] M. Flack, "Flying Stress," London: Medical Research Committee, 1918.

[45] D.J. Folds, J.M. Gerth, W.R. Engelman, " Enhancement of Human Performance in Manual Target Acquisition and Tracking, " Final Technical Report USAFASM-TR-86–18, USAF School of Aerospace Medicine, Brooks AFB, TX, 1986.

[46] D.J. Folds, " Response Organization and Time-Sharing in Dual-Task Performance," Ph.D. dissertation, School of Psychology, Georgia Institute of Technology, Atlanta, May 1987.

[47] C.R. Frankish, D.M. Jones and K.C. Kapeshi, "Maintaining Recognition Accuracy During Data Entry Tasks Using Speech Input," *Contemporary Ergonomics,* Ed E.J. Lovesey. Taylor and Francis, pp 445–453, 1990.

[48] S. Furui, "An overveiw of speaker recognition technology," *ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification,* pp. 1–9, April 1994.

[49] A.W.K. Gaillard and C.J.E. Wientjes, "Mental Load and Work Stress as Two Types of Energy Mobilization," *Work and Stress,* No. 8, pp. 141–152, 1994.

[50] A. Gallardo-Antoln, I. Mayoral and J.M. Pardo, "Automatic Speech Recognition Under Stress Conditions", *Research Report GTH-DIE-ETSIT-UPM 2/97, Grupo De Technologa Del Habla,* Departamento De Ingeniera Electrnica, Universidad Politcnica De Madrid, Spain, Nov. 1997.

[51] M.B. Gardner, "Effect of Noise System Gain, and Assigned Task on Talking Levels in Loudspeaker Communication," *J. Acoust. Soc. Am.,* **40**:955–965, 1966.

[52] K. Gopalan, "Amplitude and Frequency Modulation Characteristics of Stressed Speech," Final Report for Summer Faculty Research Program, sponsored by AFOSR, U.S. Air Force Research Laboratory [AFRL], July 1998.

[53] C.N. Hanley, D.G. Harvey, "Quantifying the Lombard Effect," *J. of Hearing & Speech Disorders,* **30**:274–7, Aug. 1965.

[54] J.H.L. Hansen, "Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition," Ph.D. Thesis, Georgia Inst. of Tech., Atlanta, GA, 428 pgs., July 1988.

[55] J.H.L. Hansen, "Evaluation of Acoustic Correlates of Speech Under Stress for Robust Speech Recognition." *IEEE Proc. 15th Bioengineering Conf.,* pp. 31–32, Boston, Mass., March 1989.

[56] J.H.L. Hansen, "A New Speech Enhancement Algorithm Employing Acoustic Endpoint Detection and Morphological Based Spectral Constraints," *ICASSP-91: IEEE Proc. Inter. Conf. on Acoust., Speech, and Signal Proc.,* pp. 901–904, Toronto, Canada, May 1991.

[57] J.H.L. Hansen, "Adaptive Source Generator Compensation and Enhancement for Speech Recognition in Noisy Stressful Environments," *IEEE 1993 ICASSP,* pp. 95–98, 1993.

[58] J.H.L. Hansen, "Morphological Constrained Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect," *IEEE Trans. Speech, Audio Proc., SPECIAL ISSUE: Robust Speech Recognition,* vol. 2(4), pp. 598–614, Oct. 1994.

[59] J.H.L. Hansen, "Analysis and Compensation of Noisy Stressful Speech for Environmental Robustness in Speech Recognition," (invited tutorial), *NATO-ESCA Proc. Inter. Tutorial & Research Workshop on Speech Under Stress,* pp. 91–98, Lisbon, Portugal, Sept. 1995.

[60] J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition," *Speech Communications, Special Issue on Speech Under Stress,* vol. 20, pp. 151–173, Nov. 1996.

[61] J.H.L. Hansen, "An Analysis of Acoustic Correlates of Speech Under Stress. Part 1: Fundamental Frequency, Duration, and Intensity Effects," submitted to *Journal Acoust. Society of America,* October 1998.

[62] J.H.L. Hansen, "An Analysis of Acoustic Correlates of Speech Under Stress. Part 2: Glottal Source and Vocal Tract Spectral Effects," submitted to *Journal Acoust. Society of America,* October 1998.

[63] J.H.L. Hansen, G. Zhou, R. Sarikaya, "An Analysis of Acoustic Correlatates of Speech Under Stress. Part 3: Applications to Stress Classification and Speech Recognition," submitted to *Journal Acoust. Society of America,* April 1999.

[64] J.H.L. Hansen, S. Bou-Ghazale, "Duration and Spectral Based Stress Token Generation for Keyword Recognition Using Hidden Markov Models," *IEEE Trans. Speech and Audio Proc.,* vol. 3(5), pp. 415–421, Sept. 1995.

[65] J.H.L. Hansen and S. Bou-Ghazale, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," *EUROSPEECH-97,* Vol.4, pp. 1743–1746, Rhodes, Greece, Sept.1997.

[66] J.H.L. Hansen, O.N. Bria, "Lombard Effect Compensation for Robust Automatic Speech Recognition in Noise," *ICSLP-90: Proc. Inter. Conf. Spoken Lang. Proc.,* pp. 1125–1128, Kobe, Japan, Nov. 1990.

[67] J.H.L. Hansen, O. Bria, "Improved Automatic Speech Recognition in Noise and Lombard Effect," *EURASIP-92.* In *Signal Processing VI: Theories and Applications,* Elsevier Publishers, New York, NY, pp. 403–406, 1992.

[68] J.H.L. Hansen, D.A. Cairns, "ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments," *Speech Communications,* **16**:391–422, July 1995.

[69] J.H.L. Hansen, M. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. on Signal Processing,* vol. 39, no. 4, pp. 795–805, April 1991.

[70] J.H.L. Hansen, M.A. Clements, "Evaluation of Speech under Stress and Emotional Conditions," *Proc. Acoust. Soc. Am.,* H15, **82**(Fall Sup.):S17, Nov. 1987.

[71] J.H.L. Hansen, M.A. Clements, "Stress Compensation and Noise Reduction Algorithms for Robust Speech Recognition," *ICASSP-89: Inter. Conf. on Acoustics Speech and Signal. Proc.,* pp. 266–269, Glasgow, Scotland, May 1989.

[72] J.H.L. Hansen, M. Clements, "Source Generator Equalization and Enhancement of Spectral Properties for Robust Speech Recognition in Noise and Stress," *IEEE Trans. Speech and Audio Proc.,* vol. 3(5), pp. 407–415, Sept. 1995.

[73] J.H.L. Hansen, B.D. Womack, "Feature Analysis and Neural Network based Classification of Speech under Stress," *IEEE Trans. Speech and Audio Proc.,* vol. 4, no. 4, pp. 307–313, July 1996.

[74] B.A. Hanson, T. Applebaum, "Robust Speaker-Independent Word Recognition Using Instantaneous, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech," *IEEE 1990 ICASSP,* pp. 857–60, Apr. 1990.

[75] M.H.L. Hecker, K.N. Stevens, G. von Bismarck, C.E. Williams, "Manifestations of Task-Induced Stress in the Acoustic Speech Signal," *J. Acoust. Soc. Am.,* 44(4):993–1001, 1968.

[76] H. Hermansky, N. Morgan, H.G. Hirsch, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *IEEE 1993 ICASSP,* pp. 83–86, 1993.

[77] J. Hernando, C. Nadeu, "Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques," *ICASSP-94: IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing*, pp. 69–72, 1994.

[78] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 5, pp. 80–84, Jan. 1997.

[79] J.W. Hicks, H. Hollien, "The Reflection of Stress in Voice-1: Understanding the Basic Correlates," *1981 Carnahan Conf. on Crime Countermeasures,* 189–195, 1981.

[80] M.J. Hunt, C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," *IEEE 1989 ICASSP,* pp.262–5.

[81] H.R. Jex, "A Proposed Set of Standardized Sub-Critical Tasks For Tracking Workload Calibration," in N. Moray, *Mental Workload: Its Theory and Measurement*, New York: Plenum Press, pp. 179–188, 1979.

[82] D. Jones, "Extending the Speaker Independent ARM Continuous Speech Recognition System to Female Voices",*DRA Memo #4636,* September 1992.

[83] B.H. Juang, "Speech Recognition in Adverse Environments," *Computer, Speech & Lang.,* pp.275–94, 1991.

[84] J.C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.,* (1):510–24, 1993.

[85] J.C. Junqua, "The Influence of Acoustics on Speech Production: A Noise-induced Stress Phenomenon Known as Lombard Reflex," *Speech Communication,* vol. 20, Nos. 1–2, pp. 13–22, 1996.

[86] J.F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal," *ICASSP-90: Inter. Conf. on Acoust., Speech, Sig. Proc.,* pp. 381–384, 1990.

[87] J.F. Kaiser, "On Teager's Energy Algorithm, its Generalization to Continuous Signals," in *Proc. 4th IEEE Digital Signal Processing Workshop,* Mohonk(New Paltz), NY, Sept. 1990.

[88] J.F. Kaiser, "Some Useful Properties of Teager's Energy Operator," *ICASSP-93: Inter. Conf. on Acoust., Speech, Sig. Proc.,* Vol. 3, pp. 149–152, 1993.

[89] I. Kuroda, O. Fujiwara, N. Okamura, N. Utsuki, "Method for Determining Pilot Stress Through Analysis of Voice Communication," *Aviation, Space, & Env. Med.,* 5:528–533, 1976.

[90] J. Laver, "Monitoring Systems in the Neurolinguistic Control of Speech," *in Fromkin, V. (ed.) "Errors of Linguistic Performance." Academic Press, New York. (Re-printed in Laver, J. "The Gift of Speech", Edinburgh University Press,* 1991.

[91] H.L. Lane, B. Tranel, and C. Sisson, "Regulation of voice communication by sensory dynamics," *Journal of Acoustical Society of America,* **47**, pp. 618–624, 1970.

[92] H.L. Lane, and B. Tranel, "The Lombard Sign and the role of hearing in speech," *Journal of Speech and Hearing Research,* **14**, pp. 677–709, 1971.

[93] C. Leeks, "Operation of a speech recognizer under whole body vibration," *Royal Aircraft Establishment,* Tech Memo FDS(F)634, 1986.

[94] P. Lieberman, S. Michaels, "Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech," *J. Acoust. Soc. Am.,* **34**(7):922–7, 1962.

[95] R.P. Lippmann, M. Mack, D. Paul, "Multi-Style Training for Robust Speech Recognition Under Stress," *Proc. of the Acoustical Society of America*, 110th Meeting, QQ10, May, 1986.

[96] R.P. Lippmann, E.A. Martin and D.B. Paul, "Multi-Style Training for Robust Isolated-Word Speech Recognition," *IEEE ICASSP-87,* pp. 705–708, 1987.

[97] O. Lippold, "Physiological Tremor," *Scientific American*, vol. 224, no. 3, pp. 65–73, Mar. 1971.

[98] S. Lively, D. Pisoni, W. van Summers, R. Bernacki, "Effects of cognitive workload on speech production," *J. Acoust. Soc. Am.,* **93**(5) 2962–73, 1993.

[99] F.H. Liu, A. Acero, R.M. Stern, "Efficient joint compensation of speech for the effects of additive noise and linear filtering," *IEEE 1992 ICASSP,* pp. 257–60.

[100] E. Lombard, "Le Signe de l'Elevation de la Voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx,* **37**:101–19, 1911.

[101] F.J. Malkin, and T. Dennison, "The effect of helicopter vibration on the accuracy of a voice recognition system," *In Proceedings of National Aerospace and Electronics Conference (NAECON),* Dayton, Ohio, 1986.

[102] D. Mansour, B.H. Juang, "A Family of Distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. ASSP,* **37**:1659–71, 1988.

[103] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "Amplitude and Frequency Demodulation Using Energy Operators", *IEEE Trans. on Signal Processing,* Vol. 41, No. 4, pp. 1532–1550, Apr. 1993.

[104] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis", *IEEE Trans. on Signal Processing,* Vol. 41, No. 10, pp. 3025–3051, Oct. 1993.

[105] C. Martins, M.I. Masacrenhas, H. Meinedo, J.P. Neto. L.C. Oliveira, C. Ribeiro, I.M. Trancoso, M.C. Viana, "Spoken Language Corpora for Speech Recognition and Synthesis in European Portuguese," *Proc. of the 10th Conf. on Pattern Recog., RECPAD'98,* pp. 357–364, Lisbon, Portugal, March 1998.

[106] A. A. Minai and R. D. Williams, "Back-propagation heuristics: A study of the extended delta-bar-delta algorithm", in *IJCNN,* June 17–21, pp. 595 – 600, 1990.

[107] B.A. Mellor, R. Graham, "The Effect on Speech Recogniser Performance of Cognitive Stress Induced by a Time Constrained Task," *DRA Memo #4749,* Unreleased, Sept. 1993.

[108] T.J. Moore, Z.S. Bond, "Acoustic-phonetic changes in speech due to environmental stressors: Implications for speech recognition in the cockpit," *R.S. Jensen (Ed) Proceedings of the 4th International Symposium on Aviation Psychology,* Columbus, Ohio, 77–83, 1987.

[109] I.R. Murray, J.L. Arnott, A.F. Newell, "Hamlet - simulating emotion in synthetic speech," *Proc of Speech'88, 7th FASE symp,* Edinburgh , pp.1217–1223. 1988.

[110] I.R., Murray, "Developing HAMLET - an emotional synthetic speech system," *Invited paper, ISACC UK Newsletter,* 6, June 1989 (Inter. Soc. for Augmentative and Alternative Communication, London), pp. 4–5.

[111] I.R. Murray, "Simulating Emotion in Synthetic Speech," *Ph.D thesis,* Dundee University, October 1989.

[112] I.R. Murray, J.L. Arnott, "Evaluation of a synthetic speech system which simulates vocal emotion by rule," *Proc. of the Inst. Of Acoustics,* 12, pp. 117–123. 1990.

[113] I.R. Murray, J.L. Arnott, N. Alm, A.F. Newell, "A communication system for the disabled with emotional synthetic speech produced by rule," *Proc. Eurospeech '91, 2nd European Conf. Speech Comm. and Tech.,* pp.311–314, Genova, Italy, 24–26 Sept. 1991.

[114] I.R. Murray, J.L. Arnott, N. Alm, A.F. Newell, "Emotional synthetic speech in an integrated communication prosthesis," *Proc. 14th Annual Con. of the Rehabilitation Engineers Soc. of North America. (RESNA), Tech. for the Nineties,* Kansas City, MO, USA, 21–26 June 1991 (J.J Presberin (Ed.), pp. 311–313, RESNA Press, Washington DC, USA.

[115] I.R. Murray, J.L. Arnott, "A tool for the rapid development of new synthetic voice personalities", *Proc. ESCA Tutorial and Research Workshop, Speech and Language Technology for Disabled Persons,* Stockholm, Sweden, 31 May-2, pp. 111–114, June 1993.

[116] I.R. Murray, J.L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. of America,* 93(2), pp. 1097–1108, Feb. 1993.

[117] I.R. Murray, J.L. Arnott, E.A. Rohwer, "Modeling vocal emotion effects in synthetic speech to improve augmented communication for non-vocal people," *Institute of Acoustics Autumn Conference on Speech and Heering, Bowness-on-Windermere,* November 1994.

[118] I.R. Murray, J.L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Communication,* 16, pp. 369–390, June 1995.

[119] I.R. Murray, J.L. Arnott, "Synthesizing emotions in speech: is it time to get excited?," *Proc. ICSLP'96,* vol. 3, pp. 1816–1819, Philadelphia, October 1996.

[120] I.R. Murray, J.L. Arnott, E.A. Rohwer, "Emotional stress in synthetic speech: Progress and future directions," *Speech Communication,* Vol. 20, Nos. 1–2, Nov. 1996.

[121] I.R. Murray, C. Baber, A. J. South, "Towards a definition and working model of stress and its effects on speech," *Speech Communication,* Vol. 20, Nos. 1–2, Nov. 1996.

[122] J. Ohala, "Ethological theory and the expression of emotion in the voice," *Proc. ICSLP'96,* Philadelphia, Oct. 1996.

[123] D.B. Paul, "A Speaker-Stress Resistant HMM Isolated Word Recognizer," *IEEE 1987 ICASSP,* pp.713–716, 1987.

[124] D.B. Paul, C.J. Weinstein, R.P. Lippman, Y. Chen, "Robust HMM-Based Techniques for Recognition of Speech Produced Under Stress and in Noise," *Proc. Speechtech-86 Conf.*, pp. 241–249, April, 1986.

[125] B.L. Pellom, J.H.L. Hansen, "Automatic Segmentation and Labeling of Speech Recorded in Unknown Noisy Channel Environments," *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels,* pp. 167–170, Pont-a-Mousson, France, April 1997.

[126] B.L. Pellom, J.H.L. Hansen, "Automatic Segmentation of Speech Recorded in Unknown Noisy Channel Characteristics," *Speech Communication: Special Issue on Robust Speech Recognition in Unknown Communication Channels,* vol. 25, nos. 1–3, pp. 97–116, Aug. 1998.

[127] J.W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE,* vol. 81, pp. 1215–1247, September 1993.

[128] D.B. Pisoni, R.H. Bernacki, H.C. Nusbaum, and M. Yuchtman, "Acoustic-phonetic correlates of Speech produced in noise," *ICASSP-85: Proc. Inter. Conf. Acoust., Speech, Sig. Proc.,* pp. 1581–1584, 1985.

[129] E.C. Poulton, "Composite Model For Human Performance in Continuous Noise," *Psych. Rev,'* 86(4), pp. 361–375, 1979.

[130] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition,* Prentice Hall Signal Processing Series, Englewood Cliffs, New Jersey, 1993.

[131] P.K. Rajasekaran, G.R. Doddington, J.W. Picone, "Recognition of speech under stress and in noise," *IEEE 1986 ICASSP,* pp. 733–736, 1986.

[132] D.A. Reynolds, "Experimental evaluation of features for robust speaker recognition," *IEEE Trans. on Speech & Audio Proc.,* vol. 2, pp. 639–643, Oct. 1994.

[133] D.A. Reynolds, "The effects of telephone transmission degradations on speaker recognition performance," *IEEE 1995 ICASSP,* pp. 329–332, May 1995.

[134] D.A. Reynolds, "The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus," *IEEE 1996 ICASSP,* pp. 113–116, May 1996.

[135] A. Rosenberg, S. Parthasarathy, "Speaker Background Models for Connected Digit Password Speaker Recognition," *IEEE 1996 ICASSP,* pp. 81–84, May 1996.

[136] R. Roessler and J. W. Lester, "Vocal patterns in anxiety," Fann, Pokorny, and Williams, editors, *Phenomenology and Treatment of Anxiety.* Spectrum, New York, 1979.

[137] M. Ross, R. Duffy, H. Cooker, D. Sargeant, "Contribution of the lower audible frequencies to the recognition of emotions," *American Annals of the Deaf,* pp. 37–42, 1973.

[138] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, D. Poch, "Time and spectrum-related variabilities in stressed speech under laboratory and real conditions," *Speech Communication,* **20**:111–130, 1996.

[139] J.C. Rutledge, K.E. Cummings, D.A. Lambert, and M.A. Clements. "Synthesizing styled speech using the klatt synthesizer," *Proc. IEEE Inter. Conf. on Acoustics, Speech, and Signal Proc.*, pp. 648–651, 1995.

[140] R. Sarikaya, J.N. Gowdy, "Subband Based Classification of Speech under Stress", *IEEE 1998 ICASSP,* pp. 569–573, 1998.

[141] K. Scherer, "Adding the affective dimension: a new look in speech analysis and synthesis," *Proc. ICSLP'96,* Philadelphia, October 1996.

[142] M.J. Russell. "The development of the Speaker Independent ARM Continuous Speech Recognition System, " *RSRE Memo #4473.* Jan. 1992.

[143] K.R. Scherer, "Nonlinguistic vocal indicators of emotion and psychopathology," C. E. Izard, editor, *Emotions in Personality and Psychopathology,* pp. 493–529. Plenum, New York, 1979.

[144] Simonov, P.V., Frolov, M.V. "Analysis of the Human Voice as a Method of Controlling Emotional State: Achievements and Goals," Aviation, Space, & Environmental Sciences, Jan., 23–25, 1977.

[145] C. Sotillo, et al., "DCIEM Sleep Deprivation Study: The Design of the Map Task Dialogues," *Report DCIEM,* Doensview, Canada, 1994.

[146] B.J. Stanton, "Robust recognition of loud and Lombard speech in the fighter cockpit environment." *Ph.D. thesis,* Purdue University, 1988.

[147] B.J. Stanton, L.H. Jamieson, G.D. Allen, "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions," *IEEE 1988 ICASSP,* pp. 331–334; 1988.

[148] B.J. Stanton, L.H. Jamieson, G.D. Allen, "Robust Recognition of Loud and Lombard Speech in the Fighter Cockpit Environment," *IEEE 1989 ICASSP,* pp.675–8.

[149] . NATO RSG-10 Research Study Group on Speech, "Potentials of Speech and Language Technology Systems for Military Use: An Application and Technology-Oriented Survey," NATO: North Atlantic Treaty Organization, Defense Research Group, Technical Report AC/243 (Panel 3) TR/21, 1996.

[150] L.A. Streeter, N.H. Macdonald, W. Apple, R.M. Krauss, K.M. Galotti, "Acoustic and Perceptual Indicators of Emotional Stress," *J. Acoust. So. Am.* **73**(4) 1354–1360, 1983.

[151] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)", in *Speech Coding and Synthesis,* Edited by W. B. Kleijn and K. K. Paliwal, Elsevier Science, Amsterdam, The Netherlands, pp. 497–518, 1995.

[152] H.M. Teager, "Some Observations on Oral Air Flow During Phonation," *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, vol. ASSP–28, no. 5, pp. 599–601, Oct. 1980.

[153] H.M. Teager and S.M. Teager, "A Phenomenological Model for Vowel Production in the Vocal Tract," in *Speech Science: Recent Advances,* edited by R.G. Daniloff (College-Hill, San Diego), pp. 73–109, 1983

[154] H.M. Teager and S.M. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract," in *Speech Production and Speech Modeling,* NATO Advanced Study Institute, Vol. 55, Bonas, France, (Kluwer Academic Pub., Boston), pp. 241–261, 1989.

[155] M. Tatham, E. Lewis, E. "Prosodic assignment in SPRUCE text-to-speech synthesis," *Proc. Inst. Acoust.,* Vol. 14, No. 6, pp. 447–454, 1996.

[156] E. Uldall, "Attitudinal meanings conveyed by intonation contours," *Language and Speech,* 3, pp. 223–234, 1960.

[157] H. Wakita, "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms," *IEEE Tran. Audio & Electroacoustics,* **AU-21**:417–427, Oct. 1973.

[158] C. Willemet, C. Vloeberghs, F. Jauquet, "Influence of Stressed Speech on Speaker Recognition System." *RSG-10 Report: Study Based on the CD-ROM SUSC-0,* RMA/SIC 1997.

[159] C.E. Williams, K.N. Stevens, "On Determining the Emotional State of Pilots During Flight: An Exploratory Study," *Aerospace Medicine,* **40** 1369–1372, 1969.

[160] C.E. Williams and K.N. Stevens, "Emotions and Speech: Some Acoustical Correlates", *Journal of Acoustical Society of America,* Vol. 52, No. 4, pp. 1238–1250, 1972.

[161] B.D. Womack, "Classification and Recognition of Speech under Perceptual Stress using Neural Networks and N-D HMMs," Ph.D. Thesis, Robust Speech Processing Lab, Dept. of Electrical Engineering, Duke Univ., Dec. 1996.

[162] B.D. Womack and J.H.L. Hansen, "Classification of Speech under Stress Using Target Driven Features," *Speech Communication,* Vol. 20, Nos. 1–2, pp. 131–150, Nov. 1996.

[163] B.D. Womack, J.H.L. Hansen, "N-Channel Hidden Markov Models for Combined Stress Speech Classification and Recognition," accepted to *IEEE Trans. Speech & Audio Proc.*, Jan. 1999.

[164] W.A. Yost, *Fundamentals of Hearing,* 3rd Edition, Academic Press, San Diego, CA., pp. 153–167, 1994.

[165] G. Zhou, J.H.L. Hansen and J.F. Kaiser, "Classification of Speech under Stress Based on Features from the Nonlinear Teager Energy Operator," *ICASSP'98,* vol. 1, pp. 549–552, Seattle, WA, 1998.

[166] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "A New Nonlinear Feature for Stress Classification," *IEEE NORSIG–98*, pp. 89 – 92, 1998.

[167] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Linear and Nonlinear Speech Feature Analysis for Stress Classification," *ICSLP–98: Inter. Conf. Spoken Lang. Proc.,* vol. 3, pp. 883–886, Sydney, Australia.

[168] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress", submitted to *IEEE Trans. on Speech and Audio Processing*, Dec. 1997.

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
	RTO-TR-10 AC/323(IST)TP/5	ISBN 92-837-1027-4	UNCLASSIFIED/ UNLIMITED

5. Originator Research and Technology Organization
North Atlantic Treaty Organization
BP 25, 7 rue Ancelle, F-92201 Neuilly-sur-Seine Cedex, France

6. Title
The Impact of Speech Under "Stress" on Military Speech Technology

7. Presented at/sponsored by
the RTO Information Systems Technology Panel (IST).

8. Author(s)/Editor(s)	9. Date
Multiple	March 2000
10. Author's/Editor's Address	**11. Pages**
Multiple	112

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

Speech recognition	Stress (psychology)	Data bases
Voice communication	Human factors engineering	C3I (Command Control Communications and Intelligence)
Military operations	Command and control	
Military applications	Sleep deprivation	
Speech	Secure communication	COTS (Commercial Off-The-Shelf)
Stress (physiology)	Workloads	

14. Abstract

Military operations are often conducted under conditions of stress induced by high workload, sleep deprivation, fear and emotion, confusion due to conflicting information, psychological tension, pain, and other typical conditions encountered in the modern battlefield context. These conditions are known to affect the physical and cognitive abilities of human speech characteristics, and this study was intended to determine the actual effects of stress on voice production quality.

It is suggested that the effect of operator based stress factors on voice is likely to be detrimental to the effectiveness of communication in general, in particular to the performance of communication equipment and weapon systems equipped with vocal interfaces (e.g., advanced cockpits, command, control, and communication systems, information warfare).

Progress in the field of military based speech technology, including advances in speech based system design has been restricted due to the lack of availability of databases of speech under stress. In particular, the type of stress which an operator may experience in the modern battlefield context is not easily simulated, and therefore it is difficult to systematically collect speech data for use in research and speech system training. It is foreseen that in the future it will be necessary to improve the coordination of multi-national military forces. The need therefore exists for planned simulations with military personnel using a wide range of speech technology and addressing factors such as high workload, sleep deprivation, fear and emotion, confusion, psychological tension, pain, etc.

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION
BP 25 • 7 RUE ANCELLE
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DIFFUSION DES PUBLICATIONS
RTO NON CLASSIFIEES

L'Organisation pour la recherche et la technologie de l'OTAN (RTO), détient un stock limité de certaines de ses publications récentes, ainsi que de celles de l'ancien AGARD (Groupe consultatif pour la recherche et les réalisations aérospatiales de l'OTAN). Celles-ci pourront éventuellement être obtenues sous forme de copie papier. Pour de plus amples renseignements concernant l'achat de ces ouvrages, adressez-vous par lettre ou par télécopie à l'adresse indiquée ci-dessus. Veuillez ne pas téléphoner.

Des exemplaires supplémentaires peuvent parfois être obtenus 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 sur la liste d'envoi de l'un de ces centres.

Les publications de la RTO et de l'AGARD sont en vente auprès des agences de vente indiquées ci-dessous, sous forme de photocopie ou de microfiche. Certains originaux peuvent également être obtenus auprès de CASI.

CENTRES DE DIFFUSION NATIONAUX

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

BELGIQUE
Coordinateur RTO - VSL/RTO
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
Directeur - Recherche et développement - Communications et gestion de l'information - DRDCGI 3
Ministère de la Défense nationale
Ottawa, Ontario K1A 0K2

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

ESPAGNE
INTA (RTO/AGARD Publications)
Carretera de Torrejón a Ajalvir, Pk.4
28850 Torrejón de Ardoz - Madrid

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)
Hellenic Ministry of National Defence
Defence Industry Research & Technology General Directorate
Technological R&D Directorate
D.Soutsou 40, GR-11521, 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 documentazione tecnico-scientifica della Difesa
Via Marsala 104
00185 Roma

LUXEMBOURG
Voir Belgique

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

PAYS-BAS
NDRCC
DGM/DWOO
P.O. Box 20701
2500 ES Den Haag

POLOGNE
Chief of International Cooperation Division
Research & Development Department
218 Niepodleglosci Av.
00-911 Warsaw

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

REPUBLIQUE TCHEQUE
VTÚL a PVO Praha /
Air Force Research Institute Prague
Národní informační středisko obranného výzkumu (NISCR)
Mladoboleslavská ul., 197 06 Praha 9

ROYAUME-UNI
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

TURQUIE
Millî Savunma Başkanliği (MSB)
ARGE Dairesi Başkanliği (MSB)
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
Document Delivery
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)

Imprimé par le Groupe Communication Canada Inc.
(membre de la Corporation St-Joseph)
45, boul. Sacré-Cœur, Hull (Québec), Canada K1A 0S7