SPEECH UNDER STRESS CONDITIONS: OVERVIEW OF THE EFFECT ON SPEECH PRODUCTION AND ON SYSTEM PERFORMANCE

Herman J.M. Steeneken

TNO Human Factors Research Institute,
Soesterberg,
The Netherlands

John H.L. Hansen

Robust Speech Processing Laboratory, Duke University
Center for Spoken Language Understanding,
University of Colorado,
Boulder, Colorado, USA

ABSTRACT

The NATO research study group on "Speech and Language Technology" recently completed a three year project on the effect of "stress" on speech production and system performance. For this purpose various speech databases were collected. A definition of various states of stress and the corresponding type of stressor is proposed. Results are reported from analysis and assessment studies performed with the databases collected for this project.

1. INTRODUCTION

Stress is a psycho-physiological state characterized by subjective strain, dysfunctional physiological activity, and deterioration of performance (Gaillard, Wientjes, 1994). Stress may be induced by external factors (workload, noise, vibration, sleep loss, etc.) and by internal factors (emotion, fatigue, etc.). Among physiological consequences of stress are respiratory changes, (e.g. increased respiration rate, irregular breathing, increased muscle tension of the vocal cords, etc.). The increased muscle tension of the vocal cords and vocal tract may, directly or indirectly, adversely affect the quality of speech. Military operations are often conducted under conditions of stress, induced by physical or mental stressors, for example: high noise environments, g-force, physical workload, mental workload, sleep deprivation, fear and emotion, confusion due to conflicting information, psychological tension, pain, and other typical conditions encountered in a military work-environment. These stresses are believed to affect voice quality, and are likely to be detrimental to the performance of communication equipment (e.g. low-bit-rate secure voice systems) and systems with vocal interfaces (e.g., advanced cockpits, command, and control systems). The NATO research study group (officially AC232/IST/TG01, formerly RSG10) initiated in 1994 the project "Speech under Stress". The goals of this project were (1) to obtain reliable (objective) stress measures deduced from speech signals and (2) to study the effect of speech under stress on the performance of speech technology equipment.

Through NATO cooperation a wide international community was invited to share in the data collected in this project and to exchange experimental results. For this purpose a special workshop was organized, and a special issue in the Journal of Speech Communication was produced.

2. DEFINITIONS OF STRESS

The implication of being "under stress" is that some form of pressure is applied to the speaker, which may result in a perturbation of the speech production process, and hence of the acoustic signal. It often happens that the pressure is in some sense threatening to the speaker (especially in the context of military operations), but this is not always so, such as for workload fatigue. This reasoning necessarily implies that a "stress free" state exists, i.e. when all pressure is absent.

As a result of the study, the effect of various stressors on the layered speech production process could be identified (Murray, Baber, and South, 1996). In Table 1, a possible relation between stages in the production process and various stressors is given.

zero-order The stressors are classified according to the level at which the stressor acts. The stressors whose effects are easiest to understand are those which have a direct physical relation on the speech production process.

first-order "First-order" stressors result in physiological changes to the speech production apparatus, altering the transduction of neuromuscular commands into movement of the articulators.

second-order "Second-order" stressors are those which affect the conversion of the linguistic program 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.

third-order "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.

3. COLLECTION OF DATA BASES

For the study "Speech under Stress" various types of databases were recorded and calibrated. The idea was to cover all four stressors as described in Section 2. The databases were made widely available in order to be able to share experimental results with other researchers outside this project. The following databases were collected:

SUSC-0/1 (Speech under Stress Conditions) This database includes:

1. Recordings from fighter cockpits and controllers, stressor psychological.
2. Two recordings from fighter cockpits during realistic alert, stressor psychological (anxiety).

Table 1: Outline of speech production process , order of stressors and Taxonomy.

Main Stages of Speech Production Process	Order of Stressor	Stressor Description	Stressors
Ideation	3	Psychological	Emotion, workload, anxiety
Generation articulatory targets	2	Perceptual	Noise (Lombard), speech quality.
Muscular commands and actions	1	Physiological	Medicines, Narcotics, Fatigue, Illness, etc.
Acoustic output	0	Physical	Vibration, Acceleration, physical work load

3. Read speech sentences, stressor physiological (physical exertion).

SUSAS (Speech Under Simulated and Actual Stress) This database includes:

1. Talking styles, stressor psychological simulated,
2. Lombard speech, stressor perceptual,
3. Computer tracking tasks, stressor psychological (time pressure, workload),
4. Roller Coaster rides, stressor physical and psychological (acceleration, and exhilaration),
5. Helicopter commands and spontaneous phrases, stressor physical, perceptual, and psychological (noise, vibration, anxiety).

DLP (DERA license plate) This database consists of prompted phrases of British car numberplates, stressor psychological (time pressure).

A full description of these databases and audio examples can be obtained at the NATO Stress Web page[1], which also includes access to additional documentation.

4. ANALYSIS, CLASSIFICATION AND DETECTION OF STRESS

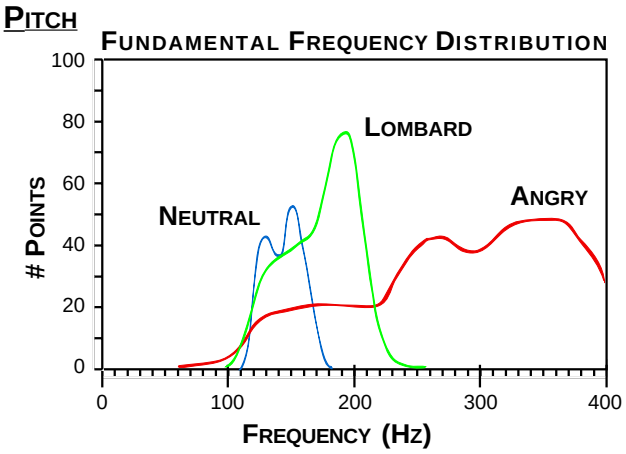
Many studies were performed on changes in the characteristics of speech produced under influence of perceptual and physical stress (noise, vibration, and g-force). However, for other types of stress research efforts were restricted to a limited number of subjects or an experimental design which was focused on a specific stressor. The parameters generally considered in evaluating changes in speech characteristics are: intensity, pitch, duration, vocal tract spectrum, glottal source and vocal tract articulatory profiles. The last two parameters can not be derived directly from the speech signal but require measurements directly related to the speaker which restricts the flexibility.

Intensity In general the average intensity is observed which increases in noise (Lombard reflex), with anger or some types of high workload. It was also found that mainly vowels and semivowels show a significant increase in intensity while consonants did not.

Pitch Pitch is the most widely considered parameter of stress evaluation. Pitch contours, variance and distributions are observed. In Fig. 1 a distribution of pitch samples is given for neutral, Lombard and angry speech.

Duration Mean word duration is a significant indicator of speech in slow, clear, angry, Lombard and loud conditions. Individual phoneme class duration under many conditions is significantly different for all styles.

[1] `http://www.ee.duke.edu/Research/Speech/stress.html`



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

Figure 1: Distribution of pitch samples for normal, Lombard, and angry speech tokens.

Vocal tract Spectrum Formant location and formant bandwidth show significant changes under various types of stress conditions. An example after (Bond et al., 1989) showing the effect of noise on the first and second formant frequency is given in Fig. 2.

Stress classification and detection is used for forensic and intelligence purposes. Research is conducted on various types of classifiers such as HMM and neural nets with Cepstral-based features and the so-called Teager Energy Operator (Teager *et al.* 1989). TEO is a non-linear differential operator which detects modulations in the speech signal and further decomposes the signal into AM and FM components. Although a number of features were investigated for stress classification, there are still many issues which need further research. These include features which change with stress, and the need for neutral and stressed reference models.

5. EFFECTS ON THE PERFORMANCE OF SPEECH TECHNOLOGY SYSTEMS

With the three recorded databases, assessment experiments were performed on speech and speaker recognition. In this paper we will report brief results obtained by Gallardo-Antolin *et al.* (1997), Hansen *et al.* (1997), and Willemet *et al.* (1998).

5.1. Speech recognition with SUSAS

To illustrate the problem of speech recognition in stress and noise, a baseline speaker-dependent, 5-state, discrete-observation HMM speech recognizer (VQ-HMM) was employed on noisefree and

Table 2: Recognition performance of speaker-dependent neutral trained discrete density HMM tested with neutral and stressed type speech in noise free and noisy conditions.

	Stressful Speech Recognition (Speaker-Dependent, Discrete-Density HMM)												
Condition	*N*	*Sl*	*F*	*So*	*L*	*A*	*C*	*Q*	*C50*	*C70*	*Lom*	μ	σ
Noise-free	88.3%	60%	65%	48%	50%	20%	68%	75%	63%	63%	63%	57.5%	15.35
Noisy†	49%	45%	28%	33%	18%	15%	40%	28%	35%	33%	28%	30.3%	9.12

†Additive white Gaussian noise, $\mathrm{SNR} = +30\,\mathrm{dB}$.

Stressed Speech Styles Key:

N – neutral	So – soft	C – clearly spoken	C50 – Moderate Load Computer Task Condition
Sl – slow	L – loud	Q – question	C70 – High Load Computer Task Condition
F – fast	A – angry		Lom – Lombard effect noise condition

Table 3: Recognition performance of speaker-independent neutral trained continuous density (2 mixtures per state) HMM models tested with neutral and stressed type speech in noise-free conditions.

Models	*Stressful Speech Recognition (Speaker-Independent, Continuous-density HMM)*										
Trained with	*N*	*C*	*C50*	*C70*	*Lom*	*So*	*Q*	*L*	*Sl*	*F*	*A*
Neutral Speech	96%	95.6%	95.4%	93.3%	91.6%	90%	85.9%	83.6%	83%	79.8%	73.5%

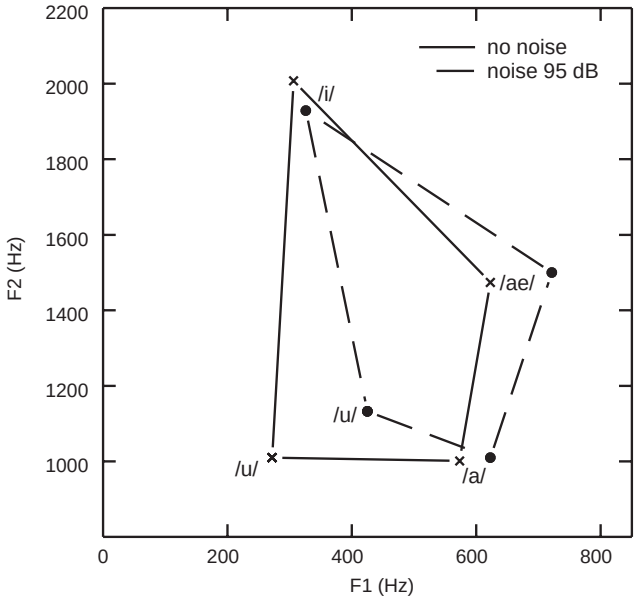


Figure 2: Shift of average center frequencies of F1 and F2 for two noise levels (male speakers).

noisy stressed speech from SUSAS (Table 2). The speaker-dependent open baseline evaluations show that when stress is present, recognition rates decrease significantly (rates for Lombard effect, loud, and angry speech range from 20–63%, with a neutral rate of 88.3%). When white Gaussian noise is introduced, noisy stressed speech rates varies, with an average rate of $Avg10 = 30.3\%$ (i.e., a 58% decrease from the 88.3% neutral rate). Recognition performance also varies considerably across stressed speaking conditions as reflected in the large standard deviation in rate of recognition ($StDev10 = 15.35$ and 9.12 for noise free and noisy stressed conditions).

In addition, the baseline recognition performance of a speaker-independent continuous density HMM employing mel-frequency cepstral parameters and time derivatives is also reported (Table 3). The data is parameterized using 12 mel-frequency cepstral coefficients. A 25 msec Hamming window is used, and a first order preemphasis is applied to the data using a coefficient of 0.97. Cepstral mean normalization is performed on the cepstral parameters to compensate for long-term spectral effects.

Table 4: Speaker-independent recognition results for a neutral trained model when tested with neutral and actual motion-fear stressed speech.

Models Trained With	**Models Tested With**	
	Neutral	Actual Stress
Neutral Speech	90.26%	26.67%

Baseline recognition performance using the fourth domain of SUSAS is reported in Table 4. A total of 35 tokens were used to train each HMM word recognizer. The speaker population employed in training the neutral models consists of 9 male speakers, while four different male speakers are employed for testing. For each word model, a total of 63 neutral tokens (7 tokens per word $\times$ 9 speakers) are employed for training a 2-mixture continuous density 5-state left-to-right HMM model. The same parameter set described previously is used for training and recognition. A total of 575 tokens is employed for testing the neutral trained speaker-independent recognizer. The recognition accuracy of the models trained and tested with neutral speech is 90.26%. Models trained with neutral speech, and tested with actual motion-fear stressed speech achieves a 26.67% recognition accuracy. Hence, the recognition error is 63.59%.

5.2. Speech recognition with DLP

The DLP data base consists of license plate numbers spoken by using the ICAO alphabet (alpha, bravo, etc.) and numbers. A total of 159 number plates spoken by 12 male and 4 female speakers was used for this assessment test. Two speaking conditions were used: a moderate dictation rate and a high dictation rate. The recognition system was an HMM based system. The training of the system was according to a round robin schedule.

5.3. Speaker recognition with SUSC0/1

For this speaker recognition experiment a system based on a vector auto-regressive method was used. Each speaker was characterized by two prediction matrices based on a second order vector-equation to predict the sequence of cepstral vectors. The system was trained with 20 s speech tokens. A test database of 10 sentences per speaker was derived from SUSC0/1. Three training-test conditions, being combinations of neutral and stressed speech, were evaluated. Also the effect of the length of the speech token

Table 5: Mean speaker-independent recognition performance (% correct) of the DLP database for two training and test conditions

Test condition	Training moderate task	Training high task
Moderate task	98	98
High task	97	98

was studied, 20 s and 40 s tokens were used. The speaker recognition performance for these conditions is given in Table 6.

Table 6: Mean speaker recognition performance for three training-test conditions and two utterance lengths.

Train-test condition	20 s utterance	40 s utterance
Neutral-neutral	95	100
Neutral-stressed	83	81
Stressed-stressed	91	99

5.4. Speaker recognition with SUSAS

A standard Gaussian mixture model was used to perform a speaker recognition task with the SUSAS database for seven speaking styles and nine male talkers. The training was performed with 35 isolated words. For each speaking style a test was performed with neutral speech data and with speech data with speaking style matched for training and test. In Table 7 the speaker recognition performance is given (% correct). Again, speaker recognition performance is significantly effected by mismatched training and test conditions

Table 7: Speaker recognition performance (% correct) with the SUSAS database for 7 speaking styles with matched and mismatched (neutral) training.

Test condition	Neutral training	Matched training
Neutral	96	96
Angry	34	75
Fast	91	90
Lombard	48	99
Loud	22	81
Slow	90	98
Soft	73	89

6. DISCUSSION AND CONCLUSIONS

The primary goal of the study reported here was to identify the effect of various types of stress on the effectiveness of communication in general, but also on the performance of communication equipment and systems equipped with vocal interfaces.

The first step, data base collection, was to make choices between realistic uncontrolled conditions or simulated (better controlled) conditions. The SUSC1/0 is an example of the first group (recording in a fighter cockpit during crash conditions). The SUSAS and DLP data bases include simulated stress by asking subjects to respond to an externally controlled condition such as speaking rate (DLP), or speaking style (SUSAS), or dual tracking computer workload (SUSAS). Finally "roller coaster" experiments simulate conditions of mental and physical stress but the subjects were not trained as a fighter pilot (SUSC0/1).

In conclusion a variety of calibrated data was collected covering a moderate range of stress conditions. Parameters indicating a change in speech characteristics as a function of the stress condition (e.g., pitch, intensity, duration, spectral envelope) were applied on several samples of stressed speech. The effect on speech obtained for perceptual (noise) and some physical stressors is evident. More difficult to determine is the effect on speech obtained for psychological and physiological stressors.

The effect of stressed speech on the performance of automatic speech recognizers and automatic speaker recognizers is for some type of stress marginal (DLP) while the speaking style has a major effect. Systems trained with matched training data (same type of speech material) do not show a major decrease in performance for stressed speech. However, for some applications obtaining such data for re-training speech systems is difficult. The project provided the NATO research study group with many interesting results but also initiated new activities. Presently a project is conducted on the "Multi-lingual Interoperability of Speech Technology Systems."

7. ACKNOWLEDGEMENTS

The work presented in this overview is the result of efforts of different Laboratories and Institutes and especially the contributions from individual researchers: Allan South DERA Farnborough UK, Roger Moore DERA Malvern UK, Carl Swail NRC-CRC Canada, Claude Vloeberghs and Patrick Verlinde RMA Belgium, Jim Cupples and John Grieco Rome Labs. USA, Isabel Trancoso INESC Portugal and José Pardo Univ. Politecnica de Madrid Spain.

8. REFERENCES

[1] Bond, Z.S., Moore, T.J., and Gable, B., (1989). "Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask." J. Acoust. Soc. Am. 85(2):907–912, 1989.

[2] Hansen, J.H.L., and Bou-Ghazale, S.E. (1997), "Getting started with SUSAS: A Speech Under Simulated and Actual Stress Database." EUROSPEECH-99, vol. 4, pp 1743–1746, Rhodes, Greece, Sept. 1997.

[3] Gaillard, A.W.K., Wientjes, C.J.E., (1994). "Mental load and work stress as two types of energy mobilization." Work and Stress, 8, 141–152.

[4] Gallardo-Antolin, A., Mayoral, I., and Pardo, J.M. (1997) "Automatic Speech Recognition under Simulated Stress Conditions." Research Report GTH-DIE-ETSIT-UPM2/97 Univ. Politecnica de Madrid.

[5] Murray, I.R., Baber, C., and South, A.J. (1996). "Towards a definition and working model of stress and its effects on Speech." Speech Communication vol. 20, Nov. 1–2, 1996.

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

[7] Willemet, C., Vloeberghs, C., and Jauquet, F. (1997). "Influence of Stressed Speech on Speaker Recognition System" RSG-10 report: study based on the CD-ROM SUSC-0, RMA/SIC Belgium.