Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering
http://pig.sagepub.com/

A speech database for stress monitoring in the cockpit

Johannes Luig and Alois Sontacchi
Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering published online 9 January 2013
DOI: 10.1177/0954410012467944

The online version of this article can be found at:
http://pig.sagepub.com/content/early/2013/01/04/0954410012467944

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Institution of MECHANICAL ENGINEERS
Institution of Mechanical Engineers

Additional services and information for *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering* can be found at:

Email Alerts: http://pig.sagepub.com/cgi/alerts

Subscriptions: http://pig.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> OnlineFirst Version of Record - Jan 9, 2013

What is This?

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

Original Article

Institution of MECHANICAL ENGINEERS

Journal of AEROSPACE ENGINEERING

A speech database for stress monitoring in the cockpit

Proc IMechE Part G:
J Aerospace Engineering
0(0) 1–13
© IMechE 2012
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0954410012467944
uk.sagepub.com/jaero
SAGE

Johannes Luig and Alois Sontacchi

Abstract

This article presents a new database of speech produced under cognitive load for the purpose of non-invasive psychological stress monitoring. The voices and the heart rates of eight airline pilots were recorded while completing an advanced flight simulation programme in a level D full flight simulator. Focusing on real-world applicability, the experiments were designed to yield the maximum degree of realism possible. Evaluation of physiological reference measures in pilots demonstrates that several heart rate variability parameters correlate with speech features derived from the recorded data. The article discusses the evolution of speech monitoring in aviation and proposes that application-orientated research methods can be useful in designing a system for real-world monitoring.

Keywords

Cockpit, database, physiological measures, pilots, speech analysis, stress monitoring

Date received: 3 January 2012; accepted: 11 October 2012

Introduction

Stress monitoring through speech analysis

Human factors in aviation are often referred to as the *bottleneck* in a highly automated environment. Since safety is of the essence, a computer algorithm cannot be the sole substitute for human decision-making and reasoning. At the same time, an individual's ability to respond is limited, and dependent on a variety of environmental factors. Only a low error rate of both air traffic control operators (ATCOs) and pilots can ensure achievement of a high level of safety. Thus, the implementation of a monitoring system that evaluates indicators of fatigue and excessive demand is reasonable. A monitoring tool could also provide air navigation supervisors with valuable insights from recent incidents, which may be used for training, process optimization, or modification of operating procedures.

For monitoring purposes, the human voice is a promising signal source for several reasons. First, it can be captured in a non-intrusive way, as the communication channel between the cockpit and the ATCO does already exist. Second, it produces a complex signal which carries a multitude of side information, so that a variety of different features can be extracted. Third, stress is a manifold phenomenon, so it is likely that different kinds and qualities of stress will be reflected in different features of the speech signal. Provided that correlations between certain value regions of speech features and manifestations of stress can be proven, a monitoring system based on speech analysis fulfils the two most important selection criteria for a measurement technique; that is, *limited intrusiveness* and *global sensitivity*.[1]

Research versus reality

Despite all the above advantages, speech monitoring methods didn't come out on top in aviation. A considerable amount of research has been done on this very topic;[2–4] generating results which look promising on the particular speech data used, but fail when applied to different datasets.[5] Section 'Fundamental considerations', takes a closer look on why existing methods fail.

In a nutshell: many studies ignore the fact that *stress* as a broadband phenomenon does not necessarily provoke a unique *stress reaction* and thus influences the process of speech production in various ways. Furthermore, the identification of critical phases requires reference speech data, which have been recorded under circumstances similar to the

Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, Austria

Corresponding author:
Johannes Luig, Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz Inffeldgasse 10/3, 8010 Graz, Austria.
Email: luig@iem.at

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

application of interest. Experiments reported in the literature have, in most cases, not been designed with the intention to use the speech data in a real-world problem.[6] Finally, and most notably, available speech databases lack additional indicators for the actual level of stress, and thus equate task complexity with the individual level of stress.

These problems are not at all aviation-specific. Schuller et al.[7] illustrate that these issues also pertain to the related area of emotion recognition in speech, an area of active research in speech processing over the last decade. Theoretical research can be focused toward robust practical application.

Section 'Database creation' presents a new database of spontaneous, unprompted aircraft pilot speech which has been recorded in a level D full flight simulator (FFL) during the completion of a custom-designed, advanced flight programme. The database is available on request (Access is subject to approval and limited to academic institutions only.) for non-commercial research purposes.

Section 'Data evaluation' evaluates the informative value of the data by investigating correlations between intended amounts of strain, physiological data, and speech parameters.

Overview: Available databases

The number of databases containing speech under stress is limited. Although Ververidis and Kotropoulos[8] list a total of 10 'emotional data collections' with cognitive stress subsets, the 1997 SUSAS database[9] is the only publicly available up-to-date reference, and thus still serves as the major source for researchers. Section 'Fundamental considerations' will, however, reveal critical issues which do not hold for this database.

Recently developed speech corpora for navigation system voice control[10–12] do fulfil requirements regarding recording environment, scope, naturalness and context (section 'General speech database requirements'), but have been created for the special case of in-vehicle human–machine communication and are not publicly available.

All other experiments reported in the literature use self-recorded, 'custom' speech data. The experimental design is not clearly stated in most cases, which influences the reproducibility of the results. Fundamental questions are left open: which part of the data was used for training, which for testing? Which evaluation strategy was chosen? Which assumptions were made when defining the *no-stress* condition?

Fundamental considerations

Stress identification through speech analysis is a classification task. Results heavily rely on (and will never be better than) those speech data taken as a reference for particular levels of stress. This section investigates how demand and response are related to each other and derive consequences for the test design.

Modeling human characteristics

A spot of system theory. On an abstract level, an individual may be modeled as a black-box system with several inputs and measureable outputs. If one provides some kind of input (question, sunshine, football), one may expect some kind of output (answer, smile, kick). The quality of output, however, shows great variation across different subjects. A human reaction may be expressed in multiple ways, and can be detected using optical, acoustical, or physiological cues (e.g. changes in heart rate and respiration, sweating, muscle tension, etc.).

A specific stimulus which is provided to provoke a reaction is not the only input to this system, as shown in Figure 1. Environmental factors need to be taken into account; as well as the intrusiveness of the measurement technique, which acts as a constant additional stressor.

Establishing direct mathematical relationships between a given input and the resulting output(s) is only allowed if the black-box system fulfils two

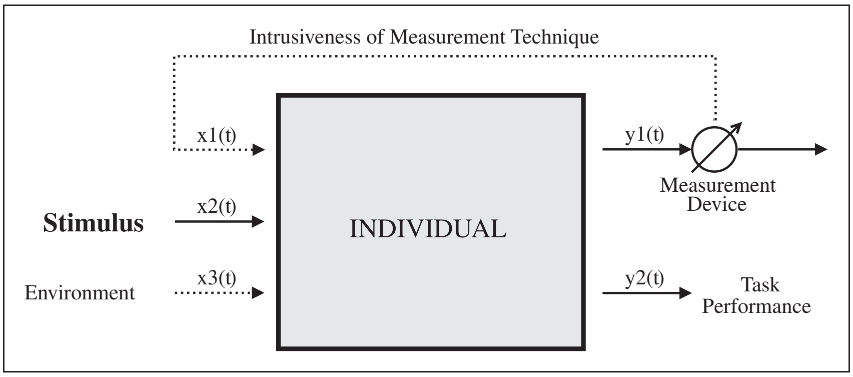

Figure 1. Technical representation of an individual as an input/output system.

important conditions; namely *linearity* and *time invariance*. Unfortunately, human beings can hardly be characterized as being linear and time-invariant in their behavior.

Demand and response capability. Stress, as psychologists define it, is 'a perceived, substantial imbalance between demand and response capability, under conditions where failure to meet the demand has important, perceived consequences'.[13] (This difference between *demand* and *subjective response capability* is constantly disregarded throughout the literature on automated speech-under-stress analysis.) An individual's ability to fulfil a given task varies depending on several parameters including age, experience, training, personality, coping behaviours, time-of-day and current mood (to name a few).

It is thus impossible to draw conclusions about individual performance from the task list, and the model's input parameters x_n are not reliable predictors for any of the outputs y_k. Apparently, the only practical way of predicting an individual's task performance ($y_2(t)$ in Figure 1) is to analyse another available output.

Environment. Influences due to the environment can be best controlled under laboratory conditions. Social influences, as the degree of communication and interaction between both pilots in the cockpit, could be controlled by reduction (single pilot operation, preassigned course of action, etc.). From an acoustical point of view, a controlled environment guarantees high-quality audio recordings free from background noise or crosstalk from another speaker. But this approach implies the risk of not being applicable in real-world scenarios, which is the case for most up-to-date available speech databases[6] (see also section 'Overview: Available databases').

Certainly, the choice of a 'realistic' environment implies limited control over crew performance and workload management.

General speech database requirements

A high-quality database of speech under stress has to meet three main requirements: scope, naturalness, and context.[14]

Scope. Scope refers to the number of speakers as well as to the number of different stress types, the number of tokens per stress type, and the gender of speakers. In most cases, available speech databases aspire to the ideal of covering the whole domain of emotions and stress, rather than focusing on a specific sub-region. This leads to a natural complexity, which is out of proportion to the number of speakers featured.

Naturalness. The majority of available speech data has not been recorded in real-life situations.[8,14] Although acted speech can be reliably classified by listeners, there are still systematic differences between acted and natural emotional speech. Acted speech is often read instead of being spoken freely, which introduces distinctive 'reading characteristics'.[15] In addition, inter-personal effects are not represented in acted speech, as it is often produced in non-interactive monologues.[14] A reference for communication in the cockpit should be generated from spontaneous, unprompted speech (which may, of course, still follow compulsory communication rules).

Context. Sample length is a crucial parameter for the applicability of speech features.[6] However, especially the widely-used database of *Speech Under Simulated and Actual Stress*[9] is limited to single-word utterances. This is a very strong constraint, since 'the way people say something' can hardly be captured without any context.

Conclusions

Additional Indicators of Stress. The authors believe that recording and analyzing speech is the least intrusive way of stress monitoring. To validate this hypothesis, reference measurements (in terms of widely accepted indicators of cognitive stress) need to be performed in addition to the speech recordings. As the choice of a 'realistic' environment limits the number of available methods for simultaneous reference parameter measurements, a mobile heart rate measurement device (section 'Measurement device') is used to ensure minimal stress impact.

Consequences for the test design. In order to meet the requirements mentioned, the experiments are conducted in a level D full flight simulator with a motion base system. The selected professional airline pilots participating in the experiments are already familiar with each other and with the simulator. The flight programme is implemented as a line-oriented flight training (LOFT) over 3.5 h to approximate the pilots' daily routine. The number of stress types is kept small in order to fulfil the criterion of limited scope. The pilots have no constraints regarding the type and amount of communication, in order to keep things as natural as possible. Finally, the context of an utterance is maintained by segmenting all speech recordings by hand.

Database creation

The concept

Physiological parameters reflect the presence of stress, and especially heart rate is useful to determine an individual's global response to task demands.[16] Laboratory experiments, including a

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

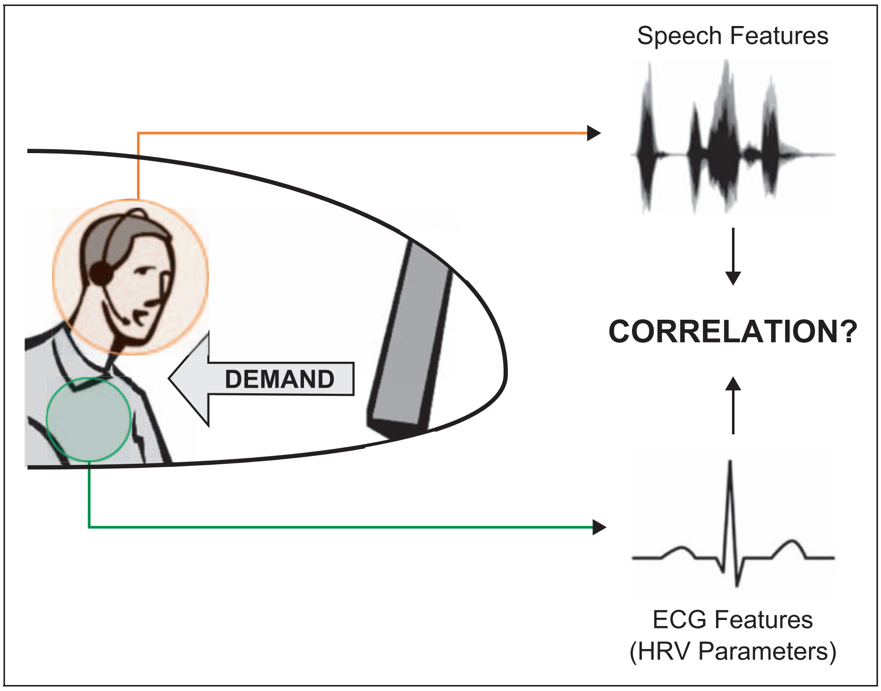

Figure 2. The speech data evaluation method using physiological measures as a reference.

recent aviation-focused study, suggest that the same is true for speech signal parameters.[2,17]

The experimental setup is aimed at exposing the pilots to a defined demand at a certain point in time, while recording both their heart rate and their speech simultaneously. The flight programme (see section 'Flight programme') was designed by professional flight instructors, who graded the expected quality and level of stress by experience. The pilots' heart rate was assessed using a mobile electrocardiogram (ECG, section 'Measurement device'); this data has afterwards been synchronized in time with the recorded speech. The research question is if correlations between physiological parameters and extracted speech features can be found in order to identify those features which indicate the presence of stress. This approach is visualized in Figure 2.

The test subjects' 'mood', which affects results without being controllable, was determined using an established questionnaire (section 'Subjective mood evaluation').

Test setup and design

Recording environment. The Fokker F70/100 series full flight simulator at Aviation Academy Austria (AAA, Neusiedl/See, Austria (www.aviationacademy.at) is a glass cockpit simulator. All instruments and controls in the cockpit are original; the motion base system simulates real flight characteristics. The simulator is equipped with high-quality vision and a 3D sound system. The audio channels (headsets and push-to-talk devices in the cockpit, flight instructor's voice) are digitally accessible in a server room outside the simulator, and can thus be captured in a completely 'invisible' way. This means that the experiments are conducted in surroundings with the highest degree of realism available on ground.

Table 1. Personal statistics of pilots participating in the recording sessions (age and experience given in years, respectively).

		Prof. experience		
Pilot ID	Age	Pilot	F100	as CMDR
CMDR 1	31	11	11	6
F/O 1	35	3	3	–
CMDR 2	45	20	20	12
F/O 2	34	12	5	–
CMDR 3	48	22	5	5
F/O 3	29	10	5	–
CMDR 4	44	16	2.5	2.5
F/O 4	29	11	4	–

CMDR: commander; F/O: first officer.

Test subjects. Eight (The number of test subjects is limited due to the fact that the reported experiments are considered a *case study*.) professional male (The overwhelming majority of available Fokker pilots were male. With respect to scope, the authors thus decided to concentrate on male test subjects first.) Fokker F70/100 pilots participated as volunteers in the experiments. The native language is (Austrian) German in seven cases; one native Danish pilot is fluent in German. The pilots were instructed to

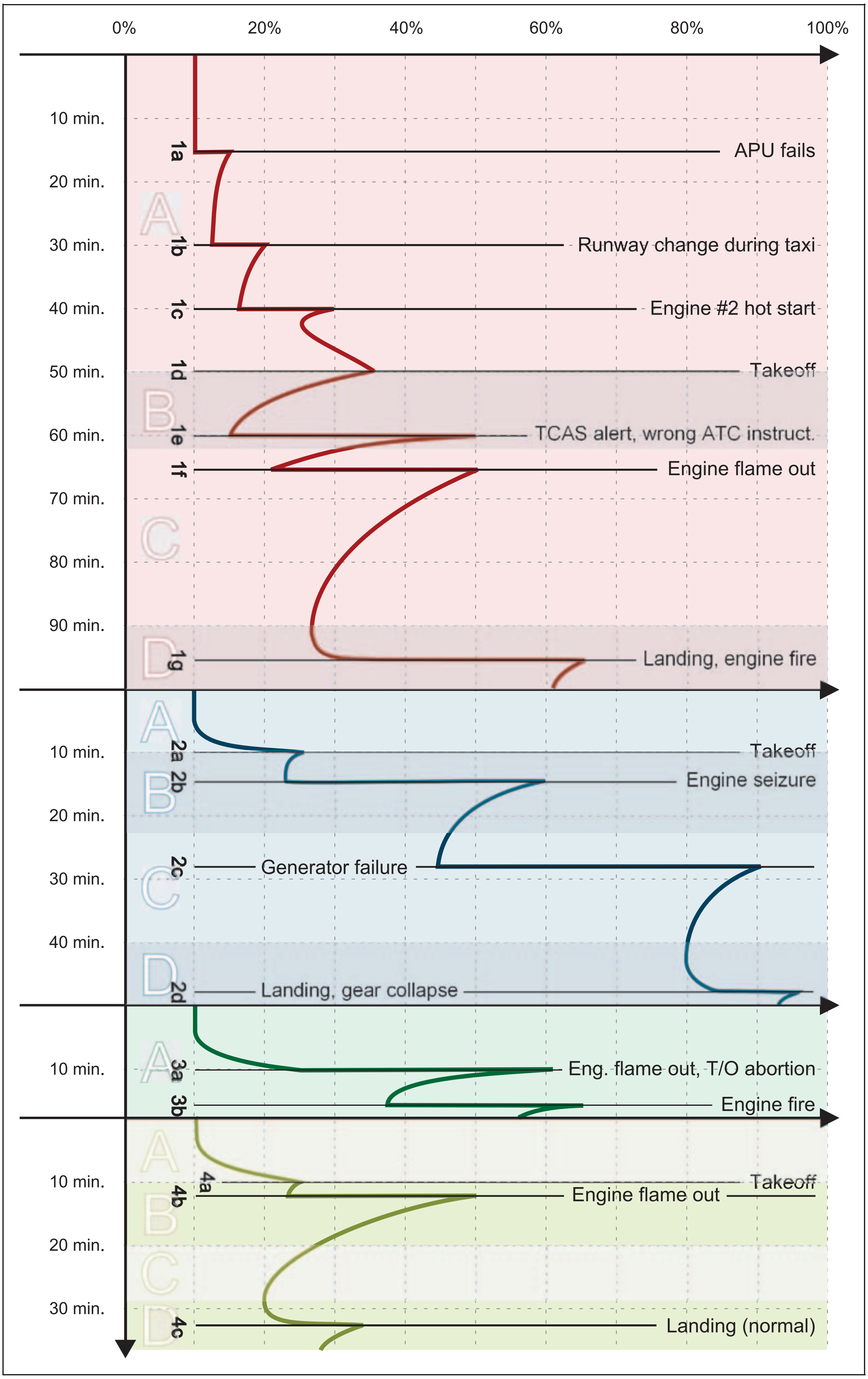

Figure 3. Strain trajectories and events/malfunctions for each of the four demanding scenarios, F1-F4. Marked flight phases are: [A] cockpit preparation, [B] takeoff and initial climb, [C] en route flight, [D] approach and landing.

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

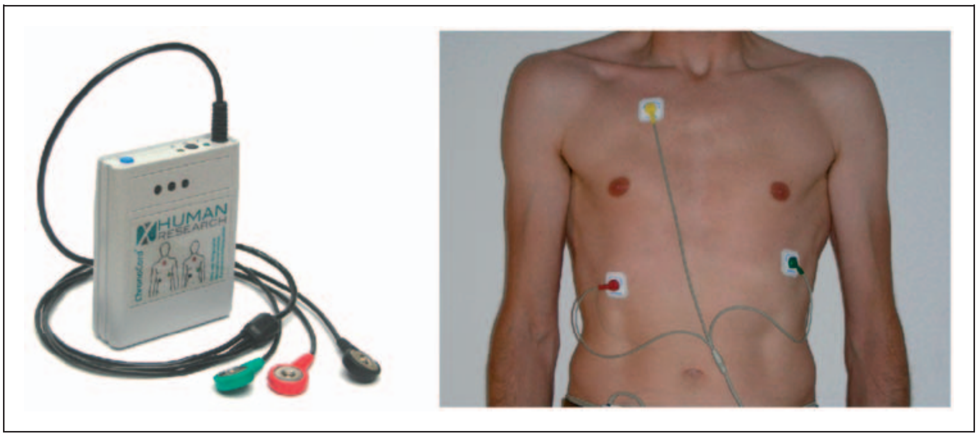

Figure 4. The ChronoCord device (left), electrode placement on a male body (right).

speak English during formal communication, while being allowed to switch to German during informal talk. The pilots were familiar with the simulator, which is used for type rating tests and biannual proficiency checks. Personal statistics of the selected volunteers are given in Table 1; the *Pilot ID* specifies session number and crew role, which is Commander (CMDR) or First Officer (F/O), respectively.

Flight programme. The recording session consisted of four different challenging scenarios (F1–F4) plus one 'reference flight' (F0) at the very beginning. The latter was included for two reasons: first, from a psychological point of view, the pilots had a warm-up flight and could acclimatize to the simulator again. Second, they provided reference values for the three main flight phases of interest; (a) takeoff and initial climb, (b) en route flight, and (c) approach and landing. The whole F0 scenario took about 20 min, and the test subjects knew in advance that nothing exceptional would be happening during this warm-up flight.

A *strain trajectory*, graded by experience, had been sketched by the instructors for each of the four demanding scenarios (Figure 3). It visualizes the presumed strain on the pilots as intended by the lesson plan; that is, the timetable of occurring events and malfunctions. The term 'strain' will be used throughout this article when referring to the demand, to emphasize the difference between task complexity and individual stress level.

Two different types of stress are included in this concept: *stress during task execution* (which is the case for most of the events) and *anticipatory anxiety* (before takeoff and landing).

These four scenarios were implemented as a LOFT, which is a 'full mission' simulation of scheduled flights. This includes a full cockpit preparation during the first scenario ('first flight of the day'), as well as intensive ATC and cabin crew communication during the flight. A 5-min rest period after the reference flight ensured that the test subjects would have had a sufficient amount of time to fully recover. The following four flights were simulated in one stretch,

1. from a temporal point of view; i.e. there were no pauses in between (except for a short break to fill out the mood questionnaires); and
2. from a local point of view; meaning that a scenario started at the same airport at which the previous one ended.

The *pilot flying* role was assigned to the CMDRs during F1 and F2 (the more demanding flights), and to the F/Os during F3 and F4.

Audio recording setup and quality. Pilots were asked to communicate via headset only and not to press the 'push-to-talk' button when using the radio transceiver or the line to the cabin. All speech signals were preamplified and dynamically compressed before being multiplexed into *ADAT* format (The ADAT signal is a multichannel optical signal transmitted via glass fiber). Usage of a state-of-the-art professional audio interface (*RME Multiface*) allowed simultaneous capture of all audio channels, while still providing the non-modified ADAT signal at the output ('feed-through'); that is, the audible signal which is fed back into the cockpit was not altered in any way. Single channels were recorded in the standard, uncompressed WAV format with a sampling rate of 44.1 kHz and a resolution of 16 bit.

A patch written in the graphical programming language *PureData (pd)* was used to record all relevant channels to the hard disk and to create speech activity information data at the same time (This perfectly synchronized speech activity log considerably facilitated subsequent data segmentation into single utterances.). The raw speech data furthermore had to be split into single files, at a maximum of 100 min in length, since the WAV file format does not allow file sizes greater than 2 GB. (For further details, the interested reader is referred to the IEM-PSD technical report.[18])

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

Table 2. MDMQ items [21].

Item	Class	Item	Class
Content	GB+	Sleepy	AT−
Rested	AT+	Good	GB+
Restless	CN−	At ease	CN+
Bad	GB−	Unhappy	GB−
Worn-out	AT−	Alert	AT+
Composed	CN+	Discontent	GB−
Tired	AT−	Tense	CN−
Great	GB+	Fresh	AT+
Uneasy	CN−	Happy	GB+
Energetic	AT+	Nervous	CN−
Uncomfortable	GB−	Exhausted	AT−
Relaxed	CN+	Calm	CN+

GB: good/bad; AT: awake/tired; CN: calm/nervous. *Plus* and *minus* signs indicate quality.

Although the speech recording quality conforms to compact disk (CD) standards, the 'perceived quality' of the speech recordings may be lowered by the fact that the pilots spoke through standard headset microphones which they were allowed to adjust at will. Common artefacts thus include sound level variability and distortion of consonants (especially plosives as 'p' and 't' or fricatives as 'f' and 's'). There is audible crosstalk (speaker B recorded through speaker A's microphone), but at a sufficiently lower level compared to the primary source (speaker A). This was considered to be tolerable, since the main objective for creating this database was the maximum degree of reality rather than making high-quality speech recordings.

Communication and speech data

Simulation of external communication. All external dialogue partners were simulated by the instructor sitting in the back of the simulator, invisible to the pilots. Also speaking through a headset microphone, his voice reached the pilots' ears through a band-limited communication channel. The instructor was asked to keep an unagitated tone when acting as an ATCO, as a technician or member of the cabin crew.

Speech categories. Depending on the actual context, the recorded speech data show large variations in terms of melody, rhythm, expressiveness and utterance length: ATC communication shows reduced prosodic intensity, as its focus is on clarity and transmission of information; whereas a discussion on what further action is required may be held in a more emotional way.

To facilitate the analysis, each single speech file is thus labeled with one of the following categories:

ATC Communication: Radio communication with ground control, radar, etc.

Cabin Address: Captain's address to passengers

Checking Procedures: One- or two-word communication (*eighty knots, my controls*)

Cockpit Communication: Instructions, questions, discussions, etc.

Free Speech: All conversation not related to controlling the aircraft or handling a situation (jokes, everyday talk)

Physiological measures

Measurement device. The test subjects' beat-to-beat heart rate signal was recorded with a high-precision (8 kHz sampling rate, 16-bit resolution.) mobile measurement device, the ChronoCord (Figure 4). All data were stored on a flash card and analysed offline. The ChronoCord offers the opportunity to set marker flags at the push of a button, so that heart rate and speech data could easily be synchronized by setting an 'acoustical marker' while pushing the button at the same time.

Data analysis. The raw data were further processed and analysed by professional physiologists (Data analysis was performed by HUMAN RESEARCH Institute, Weiz, Austria (www.humanresearch.at).). Pre-processing steps include outlier detection and correction, and data re-sampling on a regular time grid.

While heart rate itself, in general, is known to be a good indicator of how an individual reacts to a certain event in time, heart rate variability (HRV) parameters (The term *heart rate variability* here refers to the oscillation in the interval between successive R waves of the electrocardiogram (ECG).) are more meaningful as measures of cognitive load and stress.[19] HRV parameters, however, must be averaged over time windows of a certain length to achieve wide-sense stationarity. Physiologists commonly use window lengths of 2–5 min,[19] which is a very long time compared to common analysis window lengths in speech processing (which are several hundreds of milliseconds).

HRV parameters thus do not serve as indicators of stress reactions, but rather describe a steady state of autonomic balance, and the poor temporal resolution does not allow to simply 'map' the values of a certain parameter to some level of cognitive stress.

The following HRV parameters were calculated as potential stress indicators.

Variance parameters: The standard deviation of normal-to-normal intervals over 5 min of artifact-adjusted RR interval sequences (SDNN) is a widely-used statistical variability measure.

RSA-derived parameters: The heart rate increases during inspiration and decreases during expiration. This *respiratory sinus arrhythmia* (RSA) facilitates the recording of respiratory rate from the HRV signal and the calculation of several physiologically meaningful parameters such as the respiratory rate

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

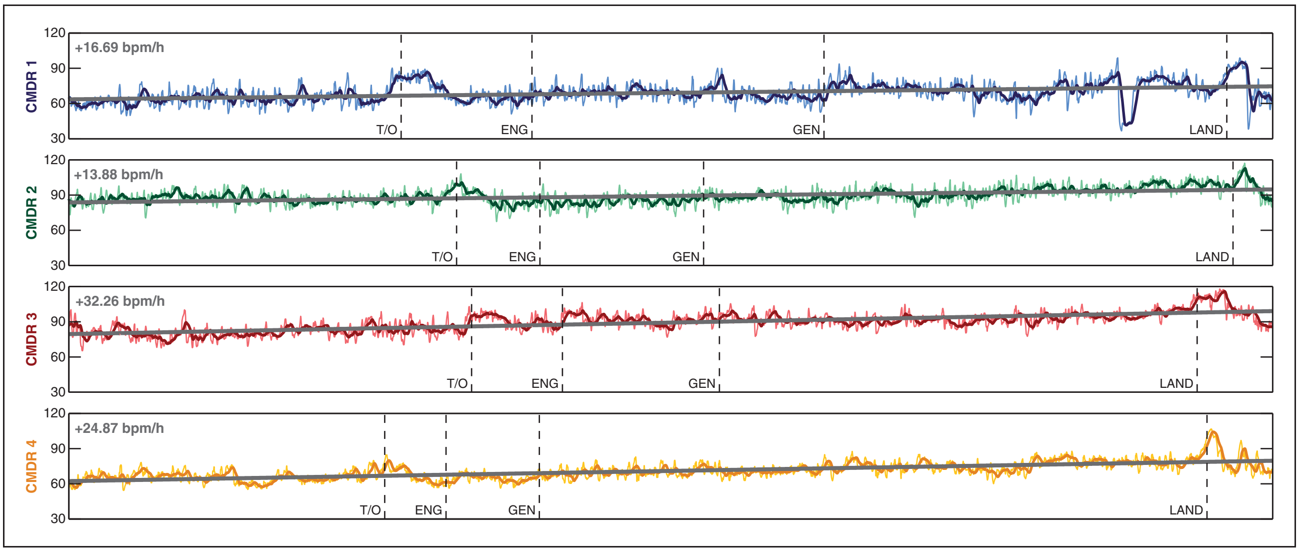

Figure 5. Comparison of the CMDRs' heart rate curves (in beats per minute, bpm) over time during F2. Fine graphs: raw heart rate signal, re-sampled at 4 Hz. Bold graphs: smoothed version (50 samples moving average). The grey lines show the global trend during F2, expressed in bpm/h. Indicated events: Takeoff, Engine Seizure, Generator Failure, Landing (with gear crash). The abscissae are normalized, i.e. the total time range shown varies over the different sessions.
CMDR: commander.

itself, the degree of modulation, and the pulse/respiration ratio.

Spectral parameters: *Power spectral density* (PSD) analysis reveals how the variability in heart rate is distributed over frequency. Two characteristic frequency bands are analysed in order to measure the balance between the sympathetic and the parasympathetic nervous system, the low-frequency band (LF, 0.04–0.15 Hz) and the high-frequency band (HF, 0.15–0.4 Hz). The LF/HF ratio is a measure for the vegetative activation level.

Rhythmic parameters: Binary encoding of the HRV derivative (i.e. $dHRV/dt$) results in patterns of zeros and ones which can be interpreted as musical rhythm patterns.[20] A distinctive predominance of certain pattern classes corresponds to a small number of different rhythmical HRV patterns and thus serves as an indicator of cardiac regularity.

For PSD analysis, an autoregressive model of order 24 was used employing the Burg algorithm, followed by integration over LF and HF bands as described by Malik et al.[19] The resulting values were compressed using the natural logarithm.

In order to remove personal characteristics such as resting heart rate, etc., the HR values were converted to *z-scores* before calculating the HRV parameters.

Subjective mood evaluation

In addition to the physiological measures, the German version of the *Multidimensional Mood State Questionnaire* (MDMQ) introduced by Steyer et al.[21] was employed to assess the current mood. The MDMQ contains 24 items in three dimensions – good/bad, awake/tired, and calm/nervous – which can be evaluated using a simple coding scheme.

This questionnaire asks the test subjects to complete the sentence *Right now, I feel…* using a 5-point scale from 1 ('definitely not') to 5 ('very much'). Table 2 gives an overview of the MDMQ items. The category acronyms, which do not appear on the questionnaires, indicate the corresponding dimension and quality of the item.

Time grids

To allow for the highest possible degree of realism, the pilots were completely free in their course of action. As a consequence, the 'unexpected' events and malfunctions occurred at different points in time (relative to the session start time). The Traffic Alert and Collision Avoidance System (TCAS), for example, should give the alarm 8 min after takeoff, during the initial climb phase. Such events were triggered by the flight instructor from the back of the cockpit.

Since these times at which the events occur are of great importance for event-based speech analysis, a detailed event log has been created manually, and the database documentation contains a list with all times of relevant events as marked in Figure 3.

Data evaluation

This section investigates correlations between the heart rate curves and the intended strain trajectories to check whether the lesson plan was succesful in provoking stress reactions from the pilots. IN a second step, both HRV and speech parameters before, during

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

and after each single event are compared to account for short- and mid-term effects.

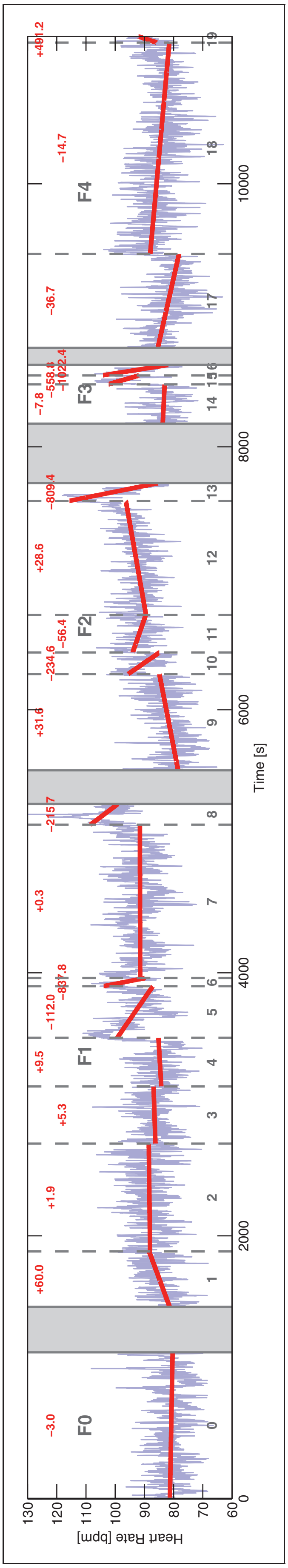

Figure 6. Heart rate trend computation for CMDR 3: raw HR curve (blue), trend lines and values (red). The reference flight F0 is included for comparison. CMDR: commander; HR: heart rate.

Heart rate and intended strain

Heart rate curves. The 'raw' heart rate signals clearly show immediate reactions to events as well as long-term trends during the sessions. Figure 5 shows the heart rates of all four commanders during the most demanding scenario, F2. The four main events during that flight are indicated by vertical lines, that is, take-off, engine flame out, generator failure, and gear crash during landing.

It is visible that characteristic signatures induced by the key events are well reproducible across subjects, at least according to the role within the crew (pilot flying vs. pilot non-flying, CMDR vs. F/O). Just as presumed, takeoff introduces an additional amount of anticipatory anxiety; the HR curves rise already before the event occurs. The general trend during F2 is also evident; to quantify this behavior, the unit *beats per minute per hour (bpm/h)* is introduced here.

Trend analysis. To compare the pilots' heart rate curves with the intended strain trajectories from Figure 3, a straight-line fit for each region between two successive events (Events 4a and 4b were combined to one single event, since the latter occurred just some 10 s after the former.) is calculated. This procedure is visualized for CMDR 3 in Figure 6.

The results, given in bpm/h, are then compared to predicted trends. To do so, the heart rate trends between consecutive events are categorized into *rising*, *falling*, and *approximately constant*. Figure 7 depicts the distribution of these categories over all regions for all pilots; the corresponding trend values are given in Table 3.

There are many reasons for the divergences from the intended categories in Figure 7. Region 1 covers about 15 min of cockpit preparation before the very first event, the differences may be caused by the pilots' current mood and expectations. In region 3, the crews show differences in coping with an announced runway change ('stay on runway' vs. 'change runway and study special performance'). Differences according to the role within the crew are visible during region 6, which implies a lot of checklist and handbook work for the F/Os, and also during region 12, where aviating the aircraft on battery power with rudimentary instruments only puts an additional strain on the CMDRs.

Interestingly, there seems to be a team component, too: tension or relaxation of one crew member might have spread to the other (as crew 4 during F2, F3, and F4).

Although about 64.5% of the trends were predicted correctly by the instructors, the considerable amount of divergence emphasizes the importance of physiological measures as a reference for the individual stress levels.

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

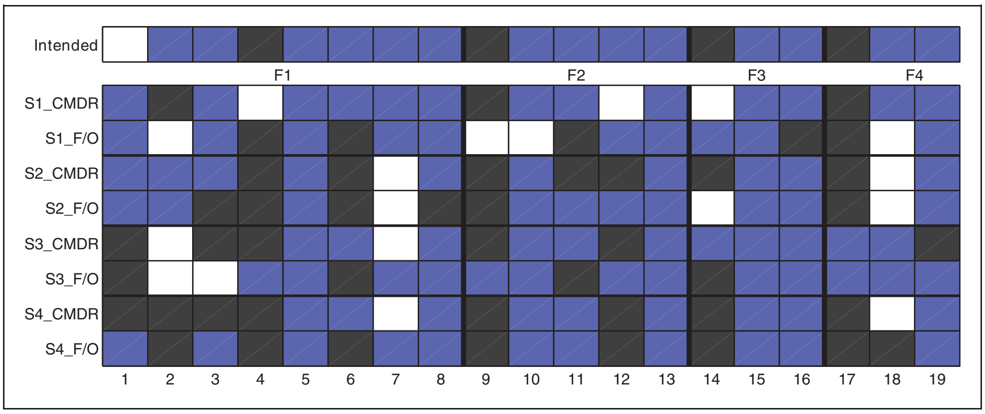

Figure 7. HR trends over all regions for all pilots and intended trend values (above). Black: rising, Blue: falling, White: approx. constant.
HR: heart rate; CMDR: commander; F/O: first officer.

Unfortunately, the MDMQ results diverge and thus are not able to provide additional evidence for global trends across all test persons.

HRV and speech parameters

HRV parameters. HRV parameters are commonly calculated using overlapping windows in the time domain with a fixed step size, resulting in a time series of parameter values.[19] This approach is not suitable in this case, as these windows cover lengths of several minutes, which results in temporal 'smearing' of the parameter values, especially if a stress reaction occurs (Figure 8, above).

Event-based analysis means that HRV parameters are calculated before, during, and after the event, using time windows of up to 5 min wherever applicable (regions A/B/C in Figure 8). The discriminative power of the HRV parameters is investigated through paired one-way ANOVA comparisons of these three regions (over all pilots), which result in 3 p-values per event.

The impact of an unexpected event may be threefold:

- There is a short-term reaction without mid-term effects; the autonomic balance is not affected. HRV parameters from regions A and C are similar, but both differ from those from region B: $(A \approx C) \neq B$.
- There is a short-term reaction which does have mid-term effects; the test person experiences another level of stress. $A \neq (B \approx C)$ or $A \neq B \neq C$.
- There is no reaction at all. $A \approx B \approx C$.

Table 3. Categories for heart rate trends.

Category	HR trend values
Rising	$> +5$ bpm/h
≈Constant	-5 bpm/h . . . $+5$ bpm/h
Falling	< -5 bpm/h

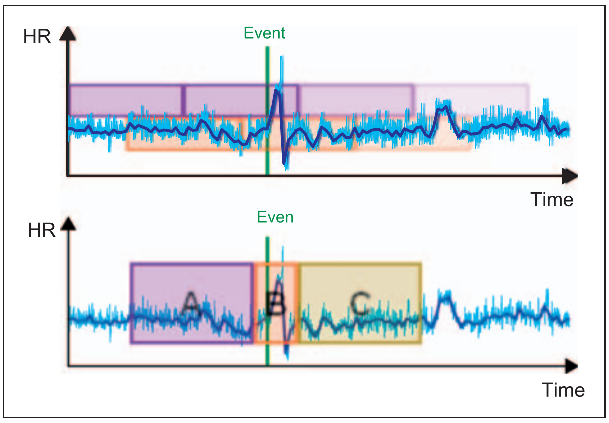

Figure 8. Common HRV parameter calculation method using overlapping windows with a fixed step size (above), event-based analysis (below).
HRV: heart rate variability.

A convenient way for compact visual interpretation is the *significance matrix*.[24] p-Values from paired comparisons are displayed in a matrix, and the interesting levels of significance, which are *significant difference* ($p < 0.05$, green) and *significant similarity* ($p > 0.95$, red), are marked using distinctive colors. Non-significant relationships are visible as grey sqares, while elements on the main diagonal represent self-comparisons, which are of no interest; they remain white.

This kind of matrix is shown for every single HRV parameter and within-flight event in Figure 9, where the column order from left to right (or, respectively, the row order from top to bottom) is before/during/after the event. One can detect characteristic patterns in these matrices. In most of the cases, all off-diagonal elements are red, indicating that HRV parameters from all three regions are significantly similar. This means, that the corresponding HRV parameter shows no sensitivity to the stress induced by the respective event.

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

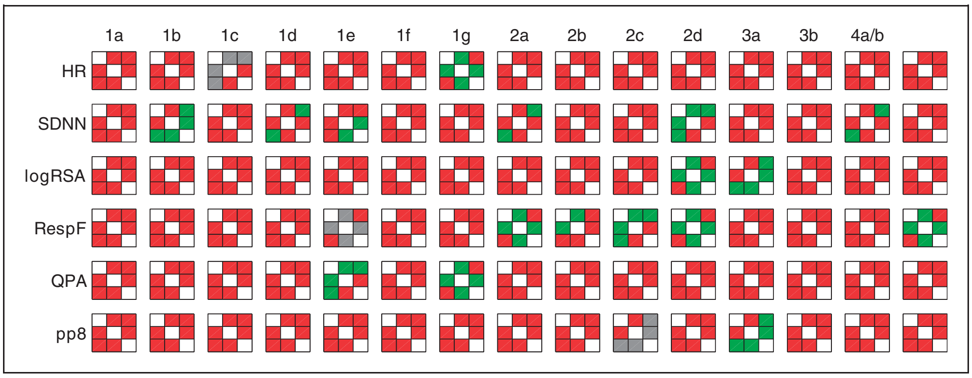

Figure 9. Significance matrices for all HRV parameters and all within-flight events. Green elements indicate significant difference ($p < 0.05$), red elements represent significant similarity ($p > 0.95$). Grey elements denote no significant relationship at all.
HRV: heart rate variability; HR: heart rate; SDNN: standard deviation of RR intervals; logRSA: median of absolute differences of successive RR intervals; RespF: respiratory rate; QPA: pulse/respiration ratio; pp8: pattern predominance.

Reactions without mid-term effects are visible as 'green crosses', e.g. HR and QPA at event 1 g. Reactions causing mid-term effects appear as a 'green gamma (Γ)', as, e.g. SDNN at event 2d. Matrices showing a 'green J', as for logRSA and pp8 at event 3a, suggest that the respective event has a delayed effect: $(A \approx B) \neq C$. Patterns showing just two green elements do not allow to derive a valid statement.

Speech parameters. For those events where at least one HRV parameter suggests significant changes in stress level, several speech parameters from the literature are extracted and correlations between HRV and speech parameter values are investigated. As mentioned in section 'Research versus rality', previous studies showed that none of those 'promising' speech parameters were able to prove universal applicability so far. As the exploration of new speech parameters for stress recognition is beyond the scope of this study, the authors stick to suggestions from the literature and extract speech parameters with prosodic relevance.

Fundamental frequency parameters: Fundamental frequency (f_0), often equated with *pitch*, is the most prominent and the most widely used speech characteristic for stress classification.[4] f_0 *mean* and f_0 *variance* are calculated over single utterances, following most approaches from the literature.

Measures of voice quality: Stevens and Hanson[22] showed that voice quality measures (VQM) can be calculated directly from the acoustic speech signal, instead of using electroglottographical techniques. Lugger et al.[23] successfully used these measures for emotion recognition under several background noise conditions. VQM parameters under investigation are *Open Quotient Gradient* (OQG), *Glottal Opening Gradient* (GOG), *Skewness Gradient* (SKG), *Rate of Closure Gradient* (RCG), and *Incompleteness of Closure* (IC).

Measures of speech tempo: Two approaches to measure speech tempo are presented here. The *syllable rate* is calculated by averaging over the first-order differences of syllable nucleus times, that is, the mean between-syllables interval. *Syllables per second* are simply the number of syllables divided by the utterance length.

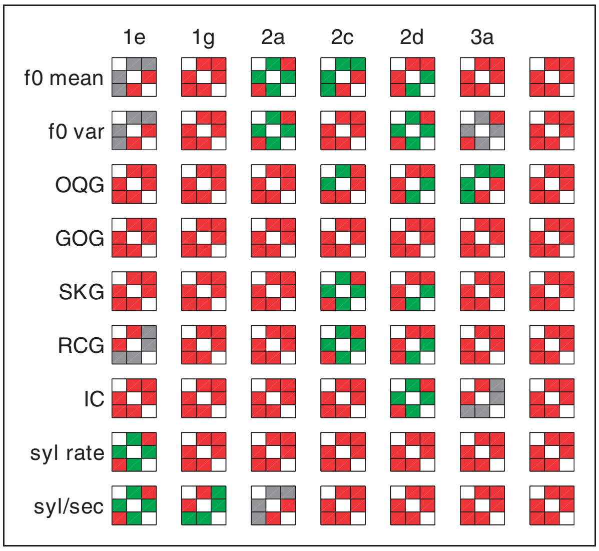

Figure 10. Significance matrices for all speech parameters and selected events. Color codes as in Figure 9.

These speech parameters are analyzed in the exact same manner as the HRV parameters described in secion 'HRV parameters'. The results, shown in Figure 10, also reveal a similar picture: there is no 'universal descriptor' for changes in stress level, but different events lead to changes in different parameters.

Interestingly, events 1e and 1 g provoke a change in speech rhythm, but leave other prosodic parameters quite unaffected. Both events (TCAS alert with wrong ATC instructions, engine fire during landing) require quick decision-making under time pressure, so it sounds reasonable that this affects the pilots' speech tempo.

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

Impact from selected F2 events (2a, 2c, 2d) is reflected in both f_0 parameters and voice quality measures, which are related due to the fact that the several gradients (OQG, GOG, SKG, RCG) express ratios of harmonics to the fundamental frequency. Just as planned, this flight turned out to be the most demanding one for all test persons, and it implicated a lot of cockpit communication to be done. In a face-to-face discussion, there is less need to control one's voice in order to be understood correctly as when, e.g. doing radio communication. This might explain the significant changes in f_0-related parameters.

Conclusion and outlook

With the creation of the IEM-PSD database, the authors have created a solid basis for research on non-invasive stress monitoring through speech analysis. To the authors' knowledge, this is the first open-source speech database of its kind. The database contains more than 7000 single utterances with a net length of more than 7 h. Eight airline pilots were recorded while completing a custom-designed, advanced flight programme in a level D full flight simulator. In addition to the speech data, each pilot's heart rate was recorded and potentially relevant HRV parameters were derived as a reference for the speaker's stress level.

The recorded heart rate data show a satisfying amount of trend correlations with the intended strain trajectories, but there is considerable divergence which emphasizes the importance of physiological measures as a reference for individual levels of stress. While the authors are convinced to have achieved the goal of creating a very realistic experimental environment (and thus bringing a remarkable degree of naturalness into the data), the experimental setup allowed great variances in the type of response to a certain event. Pilots were acting in teams without being given explicit communication rules or instructions how to manage abnormal flight conditions. The HR data do suggest the existence of team components in terms of simultaneous tension and relaxation, but differences in problem-solving between the four crews were not evaluated in detail.

In addition to the unexpected events during the flights, takeoff and landing introduce an amount of anticipatory anxiety which is not to be underestimated. While providing evidence of the pilots' 'initial conditions', the MDMQ results were unemployable for the derivation of global trends over all flights. Questionnaires measuring workload might have been more suitable for this purpose.

Paired one-way ANOVA comparisons of HRV parameters before, during, and after each single event evidence statistically significant differences in HRV parameters for nearly half of the events (7/15). This is not satisfactory and calls for a refinement of the evaluation methods, including more sophisticated techniques for compensation of personal physiological characteristics. Instead of just calculating z-scores, parameters might be computed relatively to reasonable individual reference points as the warm-up flight F0, or the 5-min rest period immediately after F0. As different flight phases (cockpit preparation, takeoff and initial climb, en route flight, approach and landing) pose different demands on the crew, one should also study common parameter responses during each phase to create global 'flight phase profiles'.

In six out of these seven cases, commonly used speech parameters support the hypotheses of significant changes in stress level. Besides, different speech parameter categories (melodic vs. rhythmic) may indicate different types of stress causes. The authors would like to emphasize the fact that this is just an intuitive selection of commonly used speech parameters to gain preliminary results. Ongoing research concentrates on the exploration of higher-level prosodic speech features beyond 'mean f_0' or 'average tempo' for the description of speech melody and rhythm.

Funding

Parts of the work reported in this paper were supported by EUROCONTROL [grant number *Graz/08-120918-C*].

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Acknowledgments

The authors thank Robert Höldrich and Peter W Pick for fruitful discussions and helpful comments. The authors are also indebted to K Wolfgang Kallus, Maximilian Moser and Nandu Goswami for good collaborations in the interdisciplinary field of stress-related research.

The main ideas of this paper have been presented at the 27th Congress of The International Council of Aeronautical Sciences (ICAS) in 2010.

References

1. Wierwille WW and Eggemeier FT. Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors* 1993; 35(2): 263–281.
2. Hansen JHL and Patil S. Speech under stress: analysis, modeling and recognition. *Speaker Classif I: Fundament Feat Meth* 2007; 1: 108.
3. Murray IR, Baber C and South A. Towards a definition and working model of stress and its effects on speech. *Speech Commun* 1996; 20(1): 3–12.
4. Hagmüller M, Rank E and Kubin G. Evaluation of the human voice for indications of workload-induced stress in the aviation environment. *EEC Note No. 18/06* 2006.
5. Luig J. *Investigations on a robust feature set for classification of speech under stress*. Master's Thesis, Graz University of Technology, 2009.

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013

6. Luig J. *Speech features for stress recognition: universal applicability of state-of-the-art methods*. Saarbrücken: VDM Verlag, 2010.
7. Schuller B, Batliner A, Steidl S, et al. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Commun* 2011; 53(9): 1062–1087.
8. Ververidis D and Kotropoulos C. Emotional speech recognition: resources, features, and methods. *Speech Commun* 2006; 48(9): 1162–1181.
9. Hansen JHL and Bou-Ghazale S. Getting started with SUSAS: a speech under simulated and actual stress database, In: *Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, vol 97, Rhodes, Greece, 22–25 September 1997, pp.1743–1746.
10. Hansen JHL, Zhang X, Akbacak M, et al. CU-Move: advanced in-vehicle speech systems for route navigation. In: Abut H, Hansen JHL and Takeda K (eds) *DSP for In-Vehicle and Mobile Systems*. New York: Springer, 2005, pp.19–45.
11. Ikeno A, Varadarajan V, Patil SA, et al. UT-Scope: speech under lombard effect and cognitive stress, In: *IEEE Aerospace Conference*, Big Sky, Montana, 3–10 March 2007, pp.1–7.
12. Boril H, Sadjadi SO and Hansen JHL. UTDrive: emotion and cognitive load classification for in-vehicle scenarios, In: *5th Biennial Workshop on DSP for In-Vehicle Systems*, Kiel, Germany, 2011.
13. McGrath JE. *Social and psychological factors in stress*. Holt: Rinehart & Winston, 1970.
14. Douglas-Cowie E, Campbell N, Cowie R, et al. Emotional speech: towards a new generation of databases. *Speech Commun* 2003; 40(1): 33–60.
15. Johns-Lewis C. Prosodic differentiation of discourse modes. *Intonation in Discourse* 1986; 199–220.
16. Wilson GF and Eggemeier FT. Psychophysiological assessment of workload in multi-task environments. In: DL Damos (ed.) *Multiple-task performance*. New York: Taylor & Francis, 1991, pp.329–360.
17. Huttunen K, Keränen H, Väyrynen E, et al. Effect of cognitive load on speech prosody in aviation: evidence from military simulator flights. *Appl Ergon* 2011; 42(2): 348–357.
18. Luig J. IEM Report 45/11: The IEM Pilot Speech Database. Technical report, Institute of Electronic Music and Acoustics, 2011.
19. Malik M, Bigger J, Camm A, et al. Guidelines: heart rate variability, standards of measurement, physiological interpretation, and clinical use. *Eur Heart J* 1996; 17: 354–381.
20. Bettermann H, Amponsah D, Cysarz D, et al. Musical rhythms in heart period dynamics: a cross-cultural and interdisciplinary approach to cardiac rhythms. *Am J Physiol-Heart Circulat Physiol* 1999; 277(5): H1762–H1770.
21. Steyer R, Schwenkmezger P, Notz P, et al. *Der Mehrdimensionale Befindlichkeitsfragebogen (in German)*. Hogrefe Verlag für Psychologie, 1997.
22. Stevens KN and Hanson HM. Classification of glottal vibration from acoustic measurements. In: O Fujimura and H Hirano (eds) *Vocal fold physiology: voice quality control*. San Diego: Singular, 1995, pp.147–170.
23. Lugger M, Yang B and Wokurek W. Robust estimation of voice quality parameters under real-world disturbances. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Toulouse, France, 14–19 May 2006, pp. I–I.
24. Frank M. *Perzeptiver Vergleich von Schallfeldreproduktionsverfahren unterschiedlicher räumlicher Bandbreite (in German)*. Master's Thesis, University of Technology, 2009.

Downloaded from pig.sagepub.com at UNIV OF TEXAS DALLAS LIBRARY on February 18, 2013