WORKLOAD MONITORING THROUGH SPEECH ANALYSIS: TOWARDS A SYSTEM FOR AIR TRAFFIC CONTROL

Johannes Luig, Alois Sontacchi
Institute of Electronic Music and Acoustics
University of Music and Performing Arts Graz, Austria

Keywords: *workload monitoring, speech analysis, feature extraction, applicability*

Abstract

Human speech is a promising signal source for workload monitoring purposes due to (a) its sensitivity to a variety of aspects of workload and (b) the facility of non-intrusive signal capturing. Many approaches in this field of research have been presented over the last years, but without leading to a working implementation in civil ATC.

In this paper, we compare proposed methods and check them for applicability in real-life scenarios. Our results show great sensitivity to the type of data investigated and demonstrate the lack of adequate ATC-related speech databases. We discuss methodic aspects and challenges regarding the creation of such a specific database. Finally, we propose a roadmap to a working system for ATC personnel workload monitoring based on speech analysis.

1 INTRODUCTION

The rapid growth of civil aviation poses a great challenge for future Air Traffic Management (ATM). Passenger safety does not only depend on the reliability of materials and instruments, but to a large extent also on *human factors*. Air Traffic Control (ATC) tasks are highly demanding in terms of situation monitoring, reasoning and decision-making.

A higher amount of air traffic will, despite all technical aid, make even higher demands on both air traffic controllers (ATCOs) and pilots, so that even greater importance will have to be attached to the balance of demand and capacity[1].

Thus, it seems reasonable to implement a monitoring system that evaluates indicators of fatigue and excessive demand in order to increase air traffic safety. Since individuals differ in physical and mental toughness, the *workload* level (describing the subjective capacity utilization) can not be derived directly from the *taskload* level (in terms of task size and complexity). What we need is a quantitative measure of instantaneous workload that facilitates automated monitoring.

Our paper is organized as follows: **section 2** is an overview of workload assessment in general and workload assessment by speech analysis in particular. In **section 3**, we present our analytical approach as well as experimental results. The findings from these results lead to a detailed discussion in **section 4**, where we highlight various aspects to consider for compiling a roadmap in **section 5**.

2 BACKGROUND

2.1 Factors affecting Workload: Stress

An individual's subjective capacity is influenced by a multitude of factors. It can be increased by, e.g., experience and training, but is certainly decreased by straining working conditions (Tab. 1).

[1]The implementation of communications, navigation, and surveillance systems in ATM (CNS/ATM) intends to reduce ATCO workload indeed, but with the goal of "increased productivity" in mind [15].

Category	Examples
Physical	Vibration, Acceleration
Physiological	Medicines, Fatigue, Illness
Perceptual	Background Noise, Poor Communication Channel
Psychological	Task-related Anxiety, Emotion

Table 1: Working conditions reducing subjective capacity.

Of course, an individual's physical and mental state does not only depend on the current working conditions, but also on long-term effects due to remarkable events and changes in private life [13].

The literature [4, 12, 10, 13, 20] subsumes the effects of the above-mentioned factors reducing subjective capacity in the term "stress"; single influences are referred to as "stressors" [12]. We will also use these terms in this paper.

2.2 The Voice as a Reference

Why do we concentrate on human speech as a signal source for workload monitoring purposes?

2.2.1 *Workload Assessment Techniques*

Throughout the literature, a combination of different techniques is used to estimate human workload during completion of a specific task: performance criteria, physiological measures, and subjective rating scales.

According to Wierwille and Eggemeier [25], the most important measurement technique selection criteria are *limited intrusiveness* – the degree of impact on task performance – and *global sensitivity*, which is the ability to discriminate between different factors affecting workload.

Different methods show significant differences in sensitivity and diagnosticity[2] [23]. Primary **task performance** alone does not allow discrimination between workload types while secondary task performance seems to reflect only major changes in workload levels [21].

[2]In this context, *diagnosticity* indicates whether a technique responds differently to different types of workload.

Concerning **physiological measures**, varying results are reported. Studies disagree about, e.g., correlation of heart rate and respiration rate or the responsibility for changes in respiration rate (metabolical vs. psychological) [2]. Furthermore, effects of work underload seem not to be reflected in physiological measures [5]. At the same time, the way of assessing the data introduces additional stress to the monitored individual; the feeling of being permanently observed is very likely to degrade performance substantially.

Finally, **subjective ratings** rely on the controllers' self-assessment and thus may provide a blurred image especially towards extreme workload levels (both low and high).

2.2.2 *Speech Signal Benefits*

The human voice, on the contrary, produces a very suitable signal for permanent monitoring purposes, since it perfectly matches the above-mentioned selection criteria:

Limited Intrusiveness – The speech signal can be measured in a non-invasive, contact-free, and non-intrusive way (Fig. 1). Data assessment is rather simple, since the required communication channel already exists, so that no additional equipment will be necessary for signal capturing. The whole recording process happens in an imperceptible way for the monitored person.

Global Sensitivity – Speech is a complex signal carrying a multitude of side information which can be accessed by extracting appropriate features (cp. section 3). The challenge is to find those properties that correlate with capacity-reducing factors (as listed in Tab. 1).

2.2.3 *Workload and Speech*

Workload effects on human speech production have been investigated during completion of specific as well as non-specific tasks.

In [16], test subjects had to perform a visual compensatory tracking task while speaking prompted sentences, which is assumed to produce cognitive load only.

The common Stroop test [28] was employed in [26] to create three defined levels of demand.

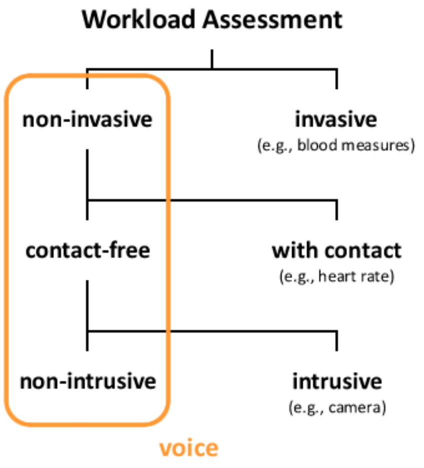

Figure 1: Properties of Workload Assessment Techniques.

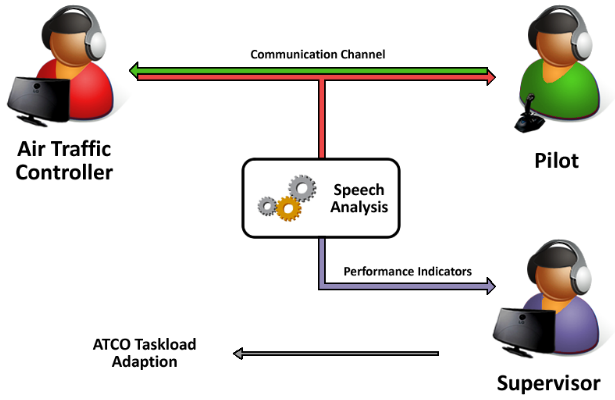

Figure 2: Outline of an ATCO Speech Monitoring System.

A dual tracking task is believed to introduce psychological stress in addition to a greater workload; experimental results [20] showed different responses for cognitive load and psychological stress.

In a task-specific experiment [7], voice analysis was performed while the test subject had to drive a car and solve mathematical questions at the same time.

2.2.4 A Speech-Based Monitoring System

Being just at the beginning of a development process of the speech monitoring system, we concentrate on the Air Traffic Controller (ATCO) first. As a matter of course, all considerations hold likewise for pilots.

The outline of our proposed speech monitoring system is depicted in Fig. 2. It analyzes the controller's voice by evaluating selected features of the speech signal which indicate different factors of human stress (cp. 2.1). The estimated workload level is directly reported to the ATCO supervisor (or, respectively, some dynamic sector sizing control software in future ATM), which in turn may adapt the demand to produce a reasonable level of workload.

3 SPEECH FEATURES

Related literature in the field [7, 10, 12, 16, 18, 20, 26, 27] is essentially concerned with discrete categories of emotions and levels of workload or stress, so speech analysis becomes a *classification* task. Available speech databases consist of short utterances, each labeled with its corresponding category as well as additional metadata including speaker identification, gender, etc.

In order to investigate the discriminative power of selected features on a specific dataset, we employ a *supervised* classification method: the algorithm is trained with one partition of the data (where class affiliation is known), before the remaining partition is used as "unknown" input to the algorithm. Performance is measured in terms of the *correct classification rate* (CCR); i.e., the ratio of correctly assigned class names to the number of samples.

3.1 Features Under Investigation

Summarizing findings from the literature, we extract the following features from the speech signal:

Fundamental Frequency (pitch) as an indicator of increased subglottal pressure in consequence of an increased respiration rate [14].

Intensity Level, reflecting an increase in vocal effort caused by a noisy environment [12] or showing larger variance due to emphasized parts of an utterance in time-critical situations [13].

Phoneme Durations, since the temporal pattern will be affected when the same amount of words is to be produced within shorter time windows between consecutive breaths [12].

Glottal Source Characteristics (in terms of jitter, shimmer, spectral level and spectral slope) in order to identify changes in the condition of the vocal folds, as a result of, e.g., a dry mouth [12].

Formant Frequencies and Bandwidths can give information about possible impact on the articulators from physiological influences like fatigue or illness [12].

Mel-Frequency Cepstral Coefficients (MFCC) represent the spectral shape in a very compact way and produce promising results, especially when using shifted delta coefficients [26].

Harmonicity as a measure of voice quality [6] and as an approximation of the signal-to-noise ratio (SNR) [1].

Zero-Crossings, indicating voiced and unvoiced speech segments.

TEO-CB-AutoEnv[3], a nonlinear feature based on the Teager Energy Operator [27].

3.2 Data Under Investigation

The widely used SUSAS database [11] contains speech under stress in five different domains; including acted emotions and speech produced during the completion of demanding computer response tasks. The vocabulary covers 35 English single-word utterances from standard aircraft communication. We take the talking styles *neutral* and *angry* as well as two computer-response tasks.

To generalize results, a German database of emotional speech (Emo-DB) [3] is used in addition. Single Emo-DB utterances are whole sentences of everyday communication, such that these data contrast with SUSAS in terms of language and sentence length. Emo-DB also contains the talking styles *neutral* and *angry*.

The ATCOSIM corpus [14] contains about 10 hours of unprompted, clean Air Traffic Control Simulation speech from non-native speakers recorded during real-time simulations. It has not been designed for the purpose of workload estimation, so that assignments of utterances to a specific workload level are missing. This restriction is overcome by taking the amount of "utterances per minute" as an indicator of the current demand and additionally weighting these values with a linearly increasing ramp to account for the fact that fatigue grows the longer a demanding task has to be performed.

3.3 Analysis Framework

The analysis framework, sketched in Fig. 3, consists of two main stages: feature extraction and feature evaluation.

At the first stage, low-level features are extracted from the buffered audio signal, resulting in one value per feature and time frame. For each of these feature series, mean and variance as well as other feature-specific characteristics are computed. Speaker dependency is eliminated in a subsequent step, before the normalized feature characteristics are evaluated regarding their ability to separate between discrete classes of emotions, workload or stress levels (depending on the data under investigation).

3.4 Experimental Design

In the first experiment, we compare optimal performing feature sets for classification of the talking styles *angry* and *neutral*, employing acted emotional speech from the SUSAS and Emo-DB databases. Classification performance is determined in terms of the *Correct Classification Rate* (CCR), which is the pooled classification result over 10 folds.

To get an idea of transferability, cross-evaluation is performed in an intermediate step: the best performing feature set for data set A (SUSAS) is applied on data set B (Emo-DB) and vice versa.

The same approach is followed for the SUSAS workload tasks (referred to as data set C and D, respectively) and the ATCOSIM speech corpus.

[3]This is the “Autocorrelation Envelope of the critical-band filtered Teager Energy Operator (TEO)”.

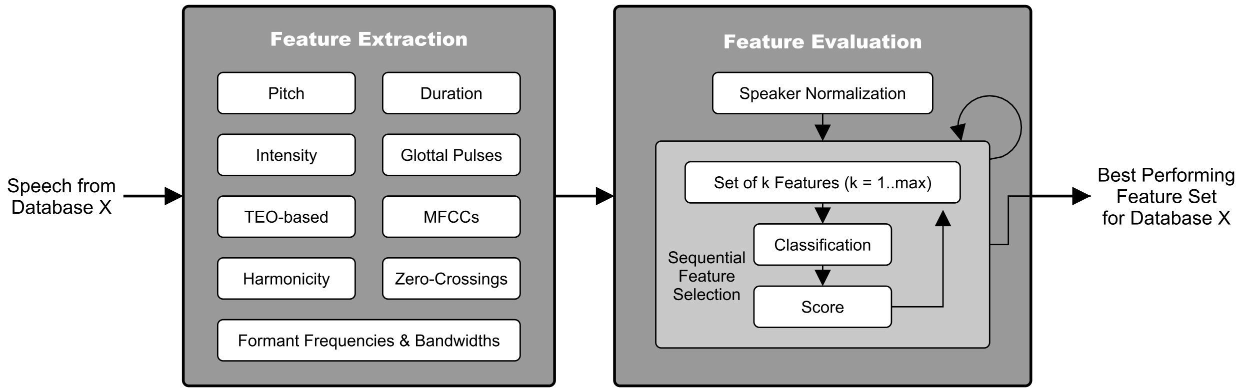

Figure 3: Feature Extraction and Evaluation Flowchart.

3.5 Results

Classification performance is satisfying concerning acted emotions. The respective optimum feature sets and corresponding classification rates are listed in Tab. 2 on page 6.

Cross-evaluation of SUSAS and Emo-DB feature sets leads to the results shown in Tab. 3: the five-feature set *F(A)* performs very well not only on the SUSAS data where it has been derived from, but also on the Emo-DB data (degrade of less than 4%).

For workload-induced stress analysis, classification results are significantly lower, as revealed in Tab. 4. The Dual Tracking task produces acceptable results; it is, however, referred to as "actual stress" in the SUSAS database (contrary to the Single Tracking task, which is labelled with "simulated stress"). Surprisingly, cross-evaluation results show that the optimal feature set found for the Dual Tracking domain performs better on the remaining two workload task domains than their "own" best performing sets.

4 DISCUSSION

4.1 Methodic Aspects and Challenges

For a working speech monitoring system as shown in Fig. 4 (on page 7), we need to define decision rules for the classification process. This includes the mapping of a certain range of feature values onto a certain region on the "degree-of-workload" scale, which is a non-trivial task. It requires

(a) knowledge of emerging stressors for the specific job, and

(b) training data containing speech produced under the influence of these stressors.

In this section, we discuss aspects to consider when it comes to the creation of such a speech database.

4.1.1 How to Simulate a Demanding Job

A working paper by Costa [4] lists the following items among the main stress sources for ATCOs[4]:

Demand – Number of aircraft under control, peak traffic hours, extraneous traffic, unforeseeable events.

Operating procedures – Time pressure, having to bend the rules, feeling of loss of control, fear of consequences of errors.

Working times – Unbroken duty periods, shift and night work

Looking at this listing, one has to question how all these stressors can be represented in a 45-minute recording session. In other words: how can we simulate effects of fatigue, monotony, shift work hours, etc. while maintaining humane work conditions? Furthermore, the paradoxical effects of work underload will certainly

[4]Costa summarizes several surveys; for details, please check the reference.

SUSAS Talking Styles	Emo-DB
Pitch: Mean Pitch: Standard Deviation Intensity: Variance Glottal Spectral Slope Harmonicity: Mean	2.MFCC TEO: Average of Means
CCR: 89.97%	CCR: 98.75%

Table 2: Best performing feature sets for acted emotions (CCR = correct classification rate).

F(A) → A	F(A) → B	F(B) → A	F(B) → B
89.97%	95.00%	68.59%	98.75%

Table 3: Cross-evaluation of feature sets for acted emotions (A = SUSAS, B = Emo-DB).

SUSAS Single Tracking	SUSAS Dual Tacking	ATCOSIM
2. MFCC Zero-Crossings: Variance	Pitch: Mean Pitch: Average Deviation Pitch: Kurtosis Intensity: Mean Glottal Spectral Mean Glottal Spectral Slope	Zero-Crossings: Variance
CCR: 56.72%	CCR: 82.70%	CCR: 63.53%

Table 4: Best performing feature sets for workload tasks.

F(C) → C	F(C) → D	F(C) → E
58.38%	55.80%	49.57%
F(D) → C	**F(D) → D**	**F(D) → E**
63.63%	82.70%	66.24%
F(E) → C	**F(E) → D**	**F(E) → E**
49.04%	53.23%	63.53 %

Table 5: Cross-evaluation of feature sets for workload tasks (C = Single Tracking, D = Dual Tracking, E = ATCOSIM).

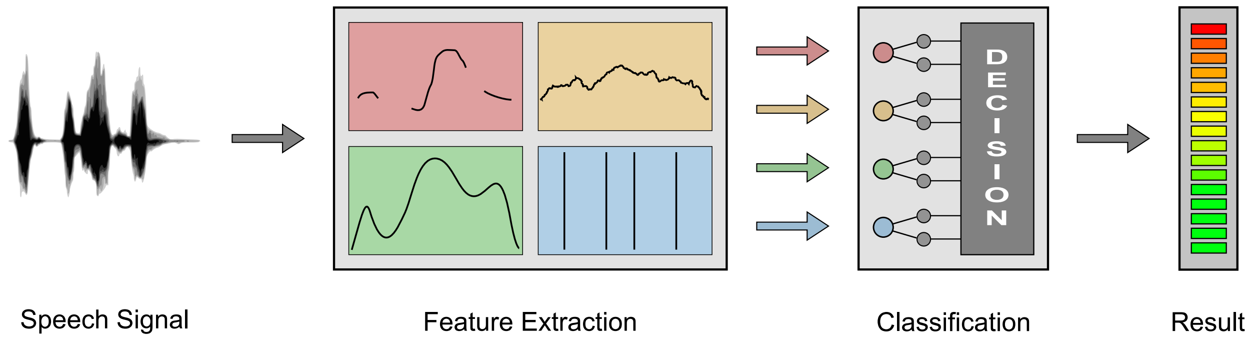

Figure 4: Working principle of proposed speech monitoring system.

appear not before a significant amount of time has passed.

Without any doubt, the ideal recording environment would be an ATCO's workplace in a "real" ATC center during regular service. Unfortunately, this option has not been existent so far due to legal reasons (data privacy) and the need for physiological reference measurements during recording (cp. 4.2.2), which may impact ATCO performance.

So the approach of choice is as follows: decompose the job into single tasks and identify impacting stressors. Design test scenarios for tracking single stress factors, do the recording and feature analysis, Finally, create a job-specific "stressor model".

This model, however, will hardly show linear behaviour, so that simple superposition of (weighted) single influences will probably not be successful. Moreover, since we assume the temporal evolution of features to be of importance, more sophisticated modeling techniques as, e.g., Hidden Markov Models (HMMs) [17] will be chosen. In this context, the question arises how much temporal information should be taken into account[5].

4.1.2 Feature Variation Between Speakers

It is a fact that phonologically identical utterances show huge acoustic variation across speakers. This *between-speaker variation* exceeds the *within-speaker variation* across emotions and stress levels in some cases; especially when analyzing pitch information, the gender of the speaker contributes in large part.

The common strategy to overcome this restriction is called *feature normalization*: by subtracting the mean and dividing by the standard deviation, all feature values are centered and scaled such that the variance equals 1. In mathematical terms,

$$f^* = \frac{f - \mu_f}{\sigma_f} \tag{1}$$

with f^* being the normalized version of feature characteristic f, while μ_f and σ_f represent the overall mean value and standard deviation of f for the current speaker.

Now that the range of values has been adjusted, an analysis-of-variance test (ANOVA) can be employed to test for statistically significant deviation of a single speaker from the pooled "group result". It is then allowed to average feature values over all individual speakers which show no significant deviation.

4.1.3 Feature Extraction and Analysis

Emotional changes happen very slowly compared to changes in the speech signal. Features are usually extracted in equidistant intervals of about $10ms$, which equals the shortest syllable duration. This degree of precision is necessary to ensure meaningful feature values, but produces a lot of single data points.

[5] A standard HMM implies the first-order Markov assumption; i.e., that the probability of a certain observation q_n at time n only depends on the observation q_{n-1} at time $n-1$.

Now, to analyze the data, it is sufficient to evaluate it in intervals of several seconds. Imagine an analysis window w of length l which is applied to the feature time series x. What we need to specify is then

(a) an adequate method to map all information within w onto a scalar value X, and

(b) a reasonable value for l.

Regarding (a), higher-order statistical moments (kurtosis) will be more succesful than simply summing up or averaging over w, since effects of transient events are preserved.

When thinking of employing prosodic features (e.g., intonation or speech rhythm) that are defined within the context of a *sentence*, another challenge is to find appropriate segmentation points. A promising approach for automatic sentence segmentation using prosodic features is presented in [8].

4.1.4 Setting the benchmark

All findings from the literature are of a *qualitative* kind so far (including ours in section 3.5). An indispensable requirement for automated analysis and decision is, however, *quantification*.

So, where to set the threshold for "capacity overload"? One possible way is to take task performance measures from simulation scenarios as a reference and to define a performance threshold which should not be fallen below.

Another important issue is the problem of the *just noticable difference* (JND). How small can a change in workload be to still be recognized; by a human supervisor on the one hand, and by an automated algorithm on the other?

Finally, the phenomenon of *work underload* has to be considered. It is frequently reported that many ATCO errors occur during periods of low demand [4]. As mentioned above, physiological measures do not seem to reflect work underload (cp. 2.2.1). Are there acoustical correlates of work underload in the speech signal?

4.1.5 Beyond "Professional Stress"

While most studies concentrate on stress introduced by straining working conditions, Hering [13] notes that "professional" and "private" stress can not be handeled separately. Remarkable events in private life or a permanent imbalance of working and leisure time increases mental load and thus reduces subjective capacity in the long run.

Since monitoring can only happen during work hours, private stress and its consequences can not be tracked and detected directly.

4.2 Speech Database Demands

4.2.1 Recording Situation

For plausible reasons, the recording environment should "look and feel" as realistic as possible.

Although the ATCO-pilot communication channel offers only poor sound quality (amplitude modulation, bandwidth $[500\ldots2500]Hz$), the audio quality should fullfil at least wideband speech standard ($[50\ldots7000]Hz$) and thus be recorded at a sampling rate of at least $16kHz$ with $16bit$ resolution. This facilitates "wideband" analysis (fundamental frequency, low energy, high-frequency consonant parts), followed by optional "narrowband" analysis of the bandpassed speech signal with additive noise in order to investigate channel effects.

Test persons should wear a headset microphone to ensure a consistent recording level.

4.2.2 Additional Measurements

Although the aim of establishing a speech monitoring system is to supersede the need for intrusive physiological measurements, it is necessary to gather additional indicators of workload as metadata. Without this reference, voice recordings could not be reconciled with any "basic truth" regarding workload. Pros and cons of physiological measurement techniques are tested and discussed in [23].

As a credible indicator of the instantaneous taskload at a point in time t, the demand (in terms of volume and complexity) has to be recorded as well.

4.2.3 Participants

The group of test subjects has to be arranged considering the following issues:

Experience of test subjects – The group should consist of experienced ATCOs as well as beginners and amateurs, assuming different amounts of subjective capacity due to different levels of experience and training.

Gender balancing – Some features show different trends in different emotional states between genders. For example, male speakers tend to lower their speech rate in anger, while female speakers speak faster in the same situation [22]. (The opposite has been noticed for sadness.)

Language – The language spoken has influence on spectral [9] and rhythmic features [19]. The latter is especially true for persons with another native language [24].

5 SUMMARY AND CONCLUSIONS

Speech analysis remains to be a promising method for workload monitoring. The functionality within specified test scenarios has been shown in the literature. Still, a wide range of open issues remains, as we have addressed in this paper.

Concluding, we propose a roadmap to a working speech monitoring system in ATC:

1. Definition of possible parameters and potential stressors for a specific ATC job.

2. Recording and analysis of speech data produced under defined influences.

3. Determination of representative speech features reflecting these influences.

4. Creation of an appropriate “job model”.

5. Implementation of speech monitoring system for the specific job model.

6. Testing under “real” conditions.

6 ACKNOWLEDGEMENTS

Work reported in this paper is kindly supported by Eurocontrol under Research Grant Scheme *Graz / 08-120918-C*. Christopher Shaw deserves to be particularly mentioned for his valuable comments and hints.

References

[1] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 17:97–110, 1993.

[2] J. Brookings, G. Wilson, and C. Swain. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, 42(3):361–377, 1996.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, and W. Sendlmeier. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology (Interspeech ’05).*, Jan 2005.

[4] G. Costa. Occupational stress and stress prevention in air traffic control. *Working Paper*, 1995.

[5] P. Desmond and T. Hoyes. Workload variation, intrinsic risk and utility in a simulated air traffic control task: evidence for compensatory effects. *Safety Science*, 22(1-3):87–101, 1996.

[6] Q. Du and C. Chang. A linear constrained distance-based discriminant analysis for hyperspectral image classification. *Pattern Recognition*, 34(2):361–373, 2001.

[7] R. Fernandez and R. Picard. Modeling drivers’ speech under stress. *Speech Communication*, Jan 2003.

[8] J. Fung, D. Hakkani-Tur, M. Doss, L. Shriberg, S. Cuendet, and N. Mirghafori. Cross-linguistic analysis of prosodic features for sentence segmentation. Technical report, Citeseer.

[9] A. Galves, J. Garcia, D. Duarte, and C. Galves. Sonority as a basis for rhythmic class discrimination. In *Speech Prosody 2002, International Conference*. ISCA, 2002.

[10] M. Hagmüller, E. Rank, and G. Kubin. Evaluation of the human voice for indications of workload-induced stress in the aviation environment. *EEC Note No. 18/06*, 2006.

[11] J. Hansen and S. Bou-Ghazale. Getting started with susas: A speech under simulated and actual stress database. *Fifth European Conference on Speech Communication and Technology (Eurospeech '97)*, Jan 1997.

[12] J. Hansen and S. Patil. Speech under stress: Analysis, modeling and recognition. *Lecture Notes in Computer Science*, Jan 2007.

[13] H. Hering. Les differents types de stress professionel appliques au controle aerien. *EEC Note No. 01/02*, 2002.

[14] K. Hofbauer, S. Petrik, and H. Hering. The atcosim corpus of non-prompted clean air traffic control speech. *spsc.tugraz.at*, 2008.

[15] ICAO. Global air navigation plan for cns/atm systems. *Doc 9750, Second Edition*, 2002.

[16] S. Lively, D. Pisoni, W. V. Summers, and R. Bernacki. Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *The Journal of the Acoustical Society of America*, Jan 1993.

[17] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[18] P. Rajasekaran, G, and J. Picone. Recognition of speech under stress and in noise. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86)*, pages 1–4, May 1986.

[19] F. Ramus. Acoustic correlates of linguistic rhythm: Perspectives. In *Speech Prosody 2002, International Conference*. Citeseer, 2002.

[20] K. Scherer, D. Grandjean, T. Johnstone, G. Klasmeyer, and T. Bänziger. Acoustic correlates of task load and stress. *Seventh International Conference on Spoken Language Processing*, 2002.

[21] N. Stanton. *Human factors methods: a practical guide for engineering and design*. Ashgate Publishing, Ltd., 2005.

[22] D. Ververidis and C. Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, Sep 2006.

[23] W. Verwey and H. Veltman. Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. *Journal of Experimental Psychology: Applied*, 2(3):270–285, 1996.

[24] L. White and S. Mattys. Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, 35(4):501–522, 2007.

[25] W. Wierwille and F. Eggemeier. Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, 35(2):263–281, 1993.

[26] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah. Speech-based cognitive load monitoring system. *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pages 2041–2044, 2008.

[27] G. Zhou, J. Hansen, and J. Kaiser. Nonlinear feature based classification of speech under stress. *Speech and Audio Processing*, Jan 2001.