ESTIMATING FATIGUE FROM PREDETERMINED SPEECH SAMPLES TRANSMITTED BY OPERATOR COMMUNICATION SYSTEMS

Jarek Krajewski[1], Udo Trutschel[2], Martin Golz[3], David Sommer[2] & Dave Edwards[4]
[1]Experimental Business Psychology, Univ. of Wuppertal, Germany;
[2]Circadian Technologies Inc., Stoneham, Massachusetts, USA;
[3]Faculty of Computer Science, University of Applied Sciences Schmalkalden, Germany;
[4]Product Safety & Compliance, Caterpillar Inc., Peoria, Illinois, USA
Email address: krajewsk@uni-wuppertal.de

Summary: We present an estimation of fatigue level within individual operators using voice analysis. One advantage of voice analysis is its utilization of already existing operator communications hardware (2-way radio). From the driver viewpoint it’s an unobtrusive, non-interfering, secondary task. The expected fatigue induced speech changes refer to the voice categories of intensity, rhythm, pause patterns, intonation, speech rate, articulation, and speech quality. Due to inter-individual differences in speech pattern we recorded speaker dependent baselines under alert conditions. Furthermore, sophisticated classification tools (e.g. Support Vector Machine, Multi-Layer Perceptron) were applied to distinguish these different fatigue clusters. To validate the voice analysis predetermined speech samples gained from a driving simulator based sleep deprivation study (N=12; 01.00-08.00 a.m.) are used. Using standard acoustic feature computation procedures we selected 1748 features and fed them into 8 machine learning methods. After each combining the output of each single classifier we yielded a recognition rate of 83.8% in classifying slight from strong fatigue.

ADVANTAGES OF SPEECH BASED FATIGUE MEASUREMENT

The working conditions of professional drivers are characterized by long working hours, movement restriction, dim light levels, background noise, and infrasonic vibration. All of these factors are known to cause fatigue, driving without awareness and even microsleep events (e.g. Horberry, Hutchins, & Tong, 2008). Thus, the prediction and warning of professional drivers against impending critical fatigue could play an important role in preventing accidents and the resulting human and financial costs. Hence, many efforts have been reported in the literature for measuring fatigue related states. But these electrode- (EOG/EEG reaching 15% error rate; Sommer & Golz, 2005) or video-based instruments (PERCLOS reaching 32% error rate;Sommer, Golz, Trutschel, & Edwards, 2008) still do not fulfill the demands of an everyday life measurement system. The major drawbacks are (a) a lack of robustness against environmental and individual-specific variations (e.g. bright light, wearing correction glasses, angle of face or being of asian race) and (b) a lack of comfort and longevity due to electrode sensor application. (cf. Figure 1).

In contrast to these electrode- or video-based instruments, the utilization of voice communication as an indicator for fatigue could match the demands of everyday life measurement. Contact free measurements such as voice analysis are non-obtrusive (not interfering with the primary driving task) and favorable for fatigue detection since an application of sensors would cause annoyance, additional stress and often impairs working capabilities and mobility demands. In addition, speech is easy to record even under extreme environmental conditions (bright light, high humidity and temperature), requires merely cheap, durable, and maintenance free sensors and most importantly, it utilizes already existing communication system hardware. Furthermore, speech data is omnipresent in many professional driver settings. Given these obvious advantages, the renewed interest in computational demanding analyses of vocal expressions has been enabled just recently by the advances in computer processing speed.

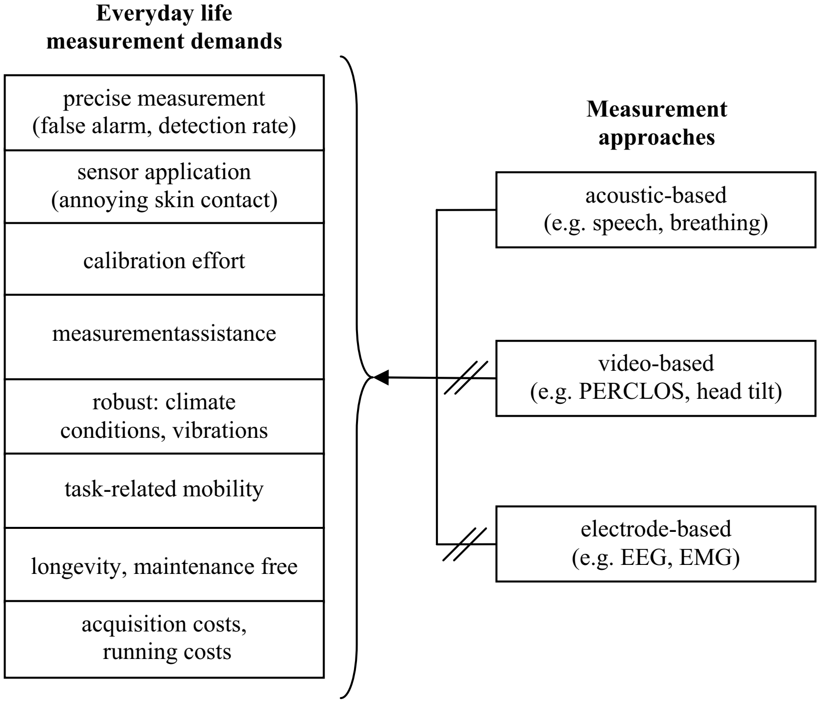
Everyday life measurement demands
precise measurement (false alarm, detection rate)
sensor application (annoying skin contact)
calibration effort
measurementassistance
robust: climate conditions, vibrations
task-related mobility
longevity, maintenance free
acquisition costs, running costs
Measurement approaches
acoustic-based (e.g. speech, breathing)
video-based (e.g. PERCLOS, head tilt)
electrode-based (e.g. EEG, EMG)

Figure 1. Demands of everyday life fatigue monitoring systems

Little empirical research has been done to examine the effect of fatigue states on acoustic voice characteristics. Previous work associating changes in voice with fatigue has generally focused on only single features (Harrison, & Horne, 1997; Whitmore, & Fisher, 1996) or small feature sets containing only perceptual acoustic features (e.g. pitch, intensity, speech rate) and using only highly artificial speech material (meaningless syllable list; Vollrath, 1994), whereas signal processing based speech and speaker recognition features (e.g. mel-frequency cepstrum coefficients, MFCCs) have received little attention (Greeley et al., 2007; Nwe, Li, & Dong, 2006). Building an automatic fatigue detection engine which reaches sufficient sensitivity and specifity still remains undone. The aim of this study is to apply a state-of-the-art speech emotion recognition engine (Batliner et al., 2006) for the detection of critical fatigue states.

The present paper will develop and evaluate a speech-based approach to estimate a speaker's level of fatigue. It is organized as follows: Section 2 introduces the cognitive-physiological mediator-model of fatigue induced speech changes. In Section 3 the procedure of computing acoustic features is explained. Section 4 describes the design of the sleep deprivation study used for building a fatigue speaker database. Having provided the results of the fatigue detection in Section 5, the paper closes with a conclusion and a discussion of the future work in Section 6.

FATIGUE INDUCED EFFECTS ON SPEECH PRODUCTION

Speech production can be described as a mental and physical process, which is realized in the following steps: create an idea, chose suitable linguistic units from memory, generate a sequence of articulatory targets, activate motor programs for the targets, transmit neuromuscular commands to muscles of the respiration and phonation system, move the respective articulators, use the proprioceptive feedback, and radiate the acoustic energy from the mouth and nostrils. Fatigue related cognitive-physiological changes - as, e.g., decreased muscle tension or reduced body temperature - can influence indirectly voice characteristics according to the following stages of speech production (Krajewski, & Kroeger, 2007): (a) Cognitive speech planning, (b) respiration, (c) phonation, (d) articulation/resonance, and (e) radiation. These changes - summarized in the cognitive-physiological mediator model fatigue induced speech changes - are based on educated guesses. In spite of the partially vague model predictions referring to fatigue sensitive acoustic features, this model provides a first insight and theoretical background for the development of acoustic measurements of fatigue (see Figure 2). Nevertheless little empirical research has been done to examine these processes mediating between fatigue, speech production, and acoustic features.

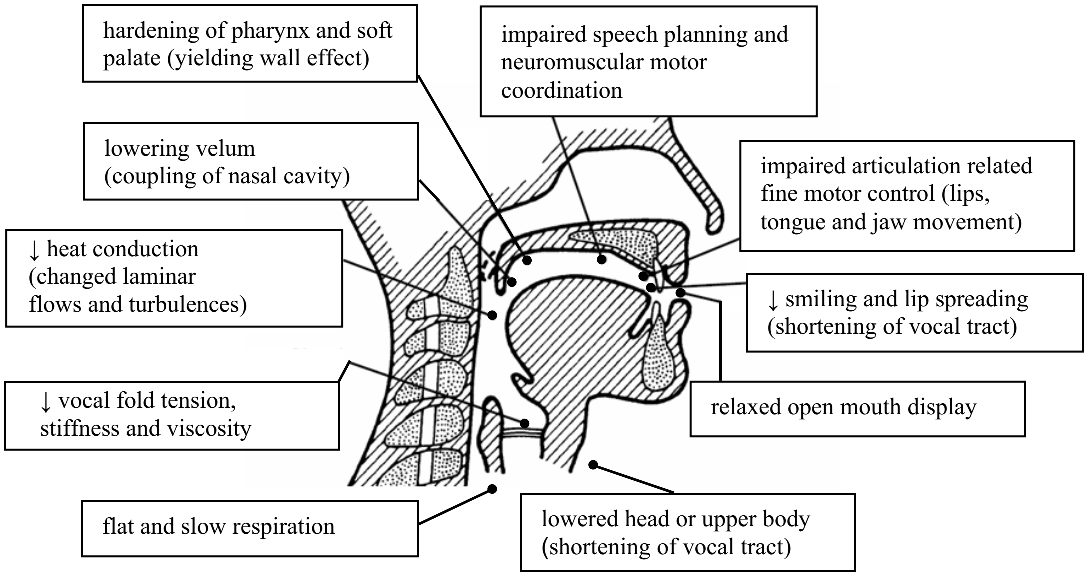
hardening of pharynx and soft palate (yielding wall effect)
impaired speech planning and neuromuscular motor coordination
lowering velum (coupling of nasal cavity)
impaired articulation related fine motor control (lips, tongue and jaw movement)
↓ heat conduction (changed laminar flows and turbulences)
↓ smiling and lip spreading (shortening of vocal tract)
↓ vocal fold tension, stiffness and viscosity
relaxed open mouth display
flat and slow respiration
lowered head or upper body (shortening of vocal tract)

Figure 2. Fatigue induced changes within the speech production system

ACOUSTIC FEATURE COMPUTATION

The acoustic fatigue analysis is mainly based on speech emotion recognition research, general audio signal and computational intelligence research (e.g. Batliner et al., 2006). Acoustic features can be divided according to auditive-perceptual concepts in prosody (pitch, intensity, rhythm, pause pattern, and speech rate), articulation (slurred speech, reduction and elision phenomena), and speech quality (breathy, whispery, tense, sharp, hoarse, or modal voice). Our approach prefers the fusion of purely signal processing based features without any known auditive-perceptual correlates and perceptual-acoustic features as e.g. the speech intensity (loudness), fundamental frequency (pitch), voiced/unvoiced duration (rhythm). Typical acoustic features used in emotion speech recognition and audio processing are (a) fundamental frequency (acoustic equivalent to pitch; rate of vocal fold vibration; models prosodic structure and speech melody), (b) intensity (models volume and stressing structure), (c) linear predictive coding coefficients, (d) harmonics-to-noise ratio (HNR; ratio between harmonic and aperiodic signal energy; breathiness and nasality indicator), (e) formant positions (resonance frequencies of the vocal tract depending strongly on lower jaw angle, tongue body angle, tongue body horizontal location, tongue tip angle, tongue tip horizontal location, relative lip height, lip protrusion, velum height), (f) formant bandwidths (model energy loss of speech signal due to vocal tract elasticity or heat conduction; yielding wall effect), (g) mel frequency cepstrum coefficients (MFCCs; separate filtering effects from the excitation signal in the time domain; entire speech production process reduced to a few decorrelated coefficients), (h) linear frequency cepstrum coefficients (LFCCs; emphasize changes or periodicity in the spectrum, while being relatively robust against noise), (i) duration of voiced/unvoiced speech segments (model temporal speech rhythm aspects as speech rate and pause structure), and (j) spectral features derived from the long term average spectrum (LTAS; relative amount of energy within predefined frequency bands; models speech quality as tense, rough and soft).

The computationally demanding feature extraction procedure results in a huge number of 45216 features and a comparatively small number of samples. This problem is well known as a curse of dimensionality and can impair the reliability of classification. Thus, the optimization of high dimensional feature spaces by means of feature selection seems a must in view of performance and real-time-capability. The acoustic measurement process follows the speech adapted steps of pattern recognition: (a) recording speech, (b) pre-processing, (c) feature computation, (d) dimensionality reduction, (e) classification, and (f) evaluation.

METHOD

Participants, Instruments and Experimental Procedure

Twelve participants took part in this study voluntarily. Initial screening excluded those having severe sleep disorders or sleep difficulties. The participants were instructed to maintain their normal sleep pattern and behaviour. Due to recording and communication problems, the data of 2 participants were only partially analyzed (2 speech samples).We conducted a within-subject sleep deprivation design (01.00 - 08.00 A.M.). During the night of sleep deprivation a well established, standardized self-report fatigue measure, the Karolinska Sleepiness Scale (KSS), was used by the participants and 2 experimental assistants nearly every hour just before the speech recordings. In the version used in the present study, scores range from ‘extremely alert’(1) to ‘extremely sleepy, can’t stay awake’ (10). Given the verbal descriptions, scores of 8 and higher appear to be most relevant from a practical perspective as they describe a state in which the subject feels unable to stay awake.

Acoustic Analysis

Recording. We conducted a validation experiment to examine whether automatically trained models can be used to recognize the fatigue of participants. Our approach can be summarized in four steps: 1. Collect individual speech data and the associated fatigue ratings for each participant; 2. Extract relevant acoustic features from the speech data; 3. Build statistical models of the fatigue ratings based on the acoustic features; 4. Test the learned models on unseen speech data. The collection of data took place in a laboratory room with dampened acoustics using a high-quality, clip-on microphone (sampling rate: 44.1 kHz, 16 bit). Furthermore, the participants were given sufficient prior practice so that they were not uncomfortable with this procedure. The verbal material consisted of predetermined formulaic operator system communication of an average length of 14 seconds. After elimination of values in the threshold region (7.5 ± 0.5 KSS) the binarization of the fatigue values leads to two classes: slight fatigue (KSS $\leq$ 7; NSS) and strong fatigue (KSS $\geq$ 8; SS). The total number of 74 speech samples was split into 27 samples of NSS and 47 samples of SS (KSS:= mean of one self-report and two observer report KSS-ratings; M= 7.47; SD= 2.59).

Feature Extraction and Pattern Recognition. To extract relevant acoustic features from the speech data all acoustic measurements were taken utterance-wise using the Praat speech analysis software for computing the acoustic features. As mentioned above, we estimated the following 59 frame-level based descriptor (FLD) contours: fundamental frequency, fundamental frequency peak process, intensity, HNR, formant position and bandwidth (F1-F6), 15 LPCs, 12 MFCCs, 12 LFCCs, 1 LFCC-flux (Euclidean distance of two consecutive LFCC frame vectors), duration of voiced, duration of unvoiced speech segments, and long term average spectrum (LTAS). In sum, we computed a total number of 45,216 features per speech sample. Further details are published elsewhere (Krajewski, Batliner, & Golz, in press). For the purpose of feature selection we used a rather relevance maximizing then redundancy minimizing correlation filter approach (Pearson correlation >.30; 1748 features remaining). To enhance robustness of the classification we feed these features separately into 8 sophisticated standard machine learning methods (multilayer perceptron (MLP), support vector machine linear kernel (SVM_l) and radial kernel (SVM_r), 5-neareast neighbour (5-NN), decision tree (DT), Parzen classifier (PC), logistic regression (LR), and linear discriminant analysis (LDA). By averaging the classification outputs, an ensemble classifier was created.

RESULTS

The spectrogram in Figure 3 provides a first insight into possible fatigue sensitive acoustic features. As we can infer from the short bright lines in the alert speaker spectrogram, the amplitudes of spectral maxima as formants and harmonics are less prominent for sleepy speakers. Furthermore, the waveform underneath documents the lower intensity variation (standard deviation of intensity, r= -.44), and a monotonous speech-pause rhythmicity in sleepy speakers (shorter voiced segment duration, r=-.41; less variation in the length of voiced segments, r=-.43). Analogous to the predictions of the cognitive-physiological mediator model proposed above, we found an energy saving flat intonation contour (1st derivate of fundamental frequency, r = -.48), and a decreased position of formant 1 values for sleepy speakers (r = -.35).

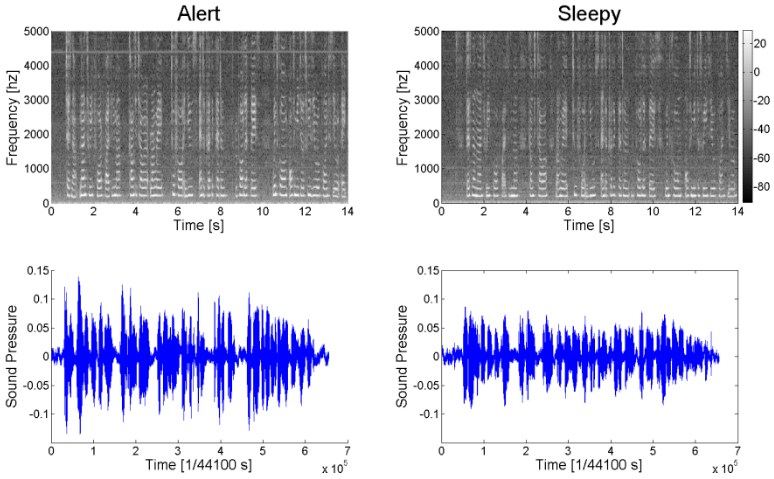

Figure 3. Spectrogram and waveform of alert vs. fatigue speech. High power spectral densities (PSD) are coded white, low PSD are coloured grey within the spectrogram.

The removal of irrelevant and redundant features often improves the performance of classification algorithms. Following the standard pattern recognition procedure, we applied a correlation-filter based feature selection method (r>.40) resulting in 1748 single acoustic features. Applying the above described ensemble classification method, this feature set achieved 83.8% recognition rate (RR; ratio correctly classified samples divided by all samples) and 82.5% class-wise averaged classification rate (CL; average of sensitivity and specificity) in the classification of sleepy speech.

DISCUSSION

The aim of the study was to construct and validate a non-obtrusive fatigue detection instrument based on predetermined speech samples as used in operator communication systems. The main findings may be summarized as following. First, acoustic features, that were extracted from speech and subsequently modelled with pattern recognition methods, contain a substantial amount of information about the speaker's fatigue state. Our acoustic measurements showed differences between alert and fatigue speech in fundamental frequency, intensity, formants, and duration features. These results are mainly consistent with the predictions of the cognitive-physiological mediator model of fatigue. Secondly, we found that a fusion of standard classifiers increases the discriminative classification power. The ensemble classifier combining 8 standard classifiers was applied on an uncommonly large feature set (45216 original acoustic features reduced to 1748 features) and yielded a recognition rate of over 83.8% on unseen data but known speakers. Our classification performance is in the same range as has been obtained for comparable tasks, e.g. for emotional user state classification (cf. Batliner et al., 2006), which are usually based on much larger databases (minimum of 200 speech samples). Thus, it seems likely to improve the fatigue detection by collecting similar sized speech databases.

There are still a number of limitations to the research presented here. The major criticism refers to the choice of the applied ground truth. The used fusion of one self-report and two observer-report measures could be criticized because of its (semi-)subjective nature lacking an involvement of "objective" physiological ground truth measures. Until now, many studies found associations between physiological data (e.g. eye blink duration) and fatigue.Nevertheless, they offer difficulties because of large inter-individual variability. This forecloses the development of commonly accepted scaling as it is realized e.g. with the KSS.

REFERENCES

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., & Aharonson, V. (2006). Combining efforts for improving automatic classification of emotional user states. In T. Erjavec, & J. Z. Gros (Eds.), *Language Technologies, IS-LTC 2006* (pp. 240-245). Ljubljana, Slovenia: Infornacijska Druzba.

Sommer, D., Golz, M., Trutschel, U, & Edwards, D. (2008). Assessing driver's hypovigilance from biosignals.*IFMBE Proceedings 22,* 152-155.

Golz, M.,& Sommer, D. (2005). Detection of strong fatigue during overnight driving. *Proceedings Annual Congress of the German Society for Biomedical Engineering, 39,* 479-480.

Greeley, H. P., Berg, J. Friets, E., Wilson, J., Greenough, G., Picone, J., Whitmore, J., & Nesthus, T. (2007). Fatigue estimation using voice analysis. *Behaviour Research Methods, 39,* 610-619.

Harrison, Y.,& Horne, J.A. (1997). Sleep deprivation affects speech. *Sleep, 20*, 871-877.

Horberry, T., Hutchins, R.,& Tong, R. (2008). Motorcycle rider fatigue: A review. *Department of TransportRoad Safety Research Report, 78,* 4-63.

Krajewski, J., Batliner, A., & Golz, M. (in press).Acoustic sleepiness detection – Framework and validation of a speech adapted pattern recognition approach. *Journal of Behavioral Research Methods.*

Krajewski, J., & Kröger, B. (2007). Using prosodic and spectral characteristics for sleepiness detection. *Interspeech Proceedings, 8,* 1841-1844.

Nwe, T.L., Li, H., &Dong, M. (2006). Analysis and detection of speech under sleep deprivation. *Interspeech Proceedings, 7,* 17-21.

Vollrath, M. (1994). Automatic measurement of aspects of speech reflecting motor coordination. *Behavior Research Methods*, *26*, 35-40.

Whitmore, J.,& Fisher, S. (1996). Speech during sustained operations. *Speech Communication, 20,* 55–70.