An Acoustic Framework for Detecting Fatigue in Speech Based Human-Computer-Interaction

Jarek Krajewski[1], Rainer Wieland[1], Anton Batliner[2]

[1] University of Wuppertal, 42097 Wuppertal, Germany
Work and Organizational Psychology
{krajewsk, wieland}@uni-wuppertal.de

[2] University of Erlangen-Nuremberg, 91058 Erlangen, Germany
Lehrstuhl fuer Mustererkennung
batliner@informatik.uni-erlangen.de

Abstract. This article describes a general framework for detecting accident-prone fatigue states based on prosody, articulation and speech quality related speech characteristics. The advantages of this real-time measurement approach are that obtaining speech data is non obtrusive, and free from sensor application and calibration efforts. The main part of the feature computation is the combination of frame level based speech features and high level contour descriptors resulting in over 8,500 features per speech sample. In general the measurement process follows the speech adapted steps of pattern recognition: (a) recording speech, (b) preprocessing (segmenting speech units of interest), (c) feature computation (using perceptual and signal processing related features, as e.g. fundamental frequency, intensity, pause patterns, formants, cepstral coefficients), (d) dimensionality reduction (filter and wrapper based feature subset selection, (un-)supervised feature transformation), (e) classification (e.g. SVM, K-NN classifier), and (f) evaluation (e.g. 10-fold cross validation). The validity of this approach is briefly discussed by summarizing the empirical results of a sleep deprivation study.

Keywords: Acoustic Features, Assistive Technologies, Pattern Recognition, Fatigue, Accident Prevention, Affective Computing

1 Measuring Fatigue in Human-Computer-Interaction

Fatigue has been widely accepted as a significant cause in a variety of traffic accidents [1-3]. Due to their slightly reduced cognitive processing speed, especially elderly persons might be vulnerable to the additional fatigue driven impairment of cognitive functions [4,5]. Similar risk factor constellations can be found in private application contexts (e.g. Telecare systems identifying critical fatigue states in senior people) as well as in general work contexts (e.g. monitoring tasks, air traffic control), too. Hence, detecting fatigue states and reacting to them is an important issue for accident prevention in elderly persons.

Side applications of detecting fatigue in Human-Computer-Interaction (HCI) can be identified within the field of Assistive Technologies for elderly and disabled persons [e.g. 6], making HCI more natural, e.g. in speech interfaces of assistive domotics. Adapting the system output to the actual emotional and fatigue related user states might enhance the acceptance of these systems due to their improved naturalism. The emotional-intelligent, user state sensitive communication could improve comfort. Furthermore it may result in better comprehensiveness if the system output is adapted to the user's actual fatigue-impaired attentional and cognitive resources. In addition to this, speech recognition or speaker verification systems itself might be improved by taking the fatigue related speech changes into account. As a results speech based Assistive Technologies could benefit from these progresses.

Many efforts have been reported in the literature for measuring biosignal based fatigue states [7]. These systems mainly focus on (a) oculomotoric data (eye blinking, eyelid movement, and saccade eye movement) [8], (b) EEG data [9] and (c) behavioral expression data (gross body movement, head movement, mannerism, and facial expression) [10] in order to characterize the fatigue and sleepiness state. Apart from these promising advances in analysing facial and gestural expressivity, there has been recently renewed interest in vocal expression and speech analysis. This fact is mainly promoted by the progress in speech science and the gaining presence of speech in voice guided HCI. Using voice communication as an indicator of sleepiness e.g. within assistive technologies for elderly would have the following advantages: obtaining speech data is non obtrusive, free from sensor application and calibration efforts, “hands- and eyes-free”, and most important, speech data is often available in HCI situations.

In this paper we describe a speech adapted pattern recognition framework in order to measure fatigue and sleepiness states. Our attention is focused particularly on the processing step of feature computation. The rest of this paper is organized as follows: In Section 2 computing high level contour descriptor features is explained. The general speech adapted pattern recognition framework is provided in Section 3, a brief summary of sleepiness detection results is given in Section 4.

2 Acoustic Features

Acoustic features can be divided according to auditive-perceptual concepts into prosody (pitch, intensity, rhythm, pause pattern, speech rate), articulation (slurred speech, reduction and elision phenomena), and speech quality (breathy, tense, sharp, hoarse, modal voice) related features [11]. Another distinction can be drawn from using signal processing categories as time, frequency or phase space domain features. Our approach prefers the fusion of perceptual features and purely signal processing and speech recognition based features without any known auditive-perceptual correlates. Typical frame level based acoustic features (Low-Level Descriptors, LLD; see [12,13]) used in emotion speech recognition and audio processing [14,15] are fundamental frequency (acoustic correlate to pitch; maximum of the autocorrelation

function), intensity, duration of voiced/unvoiced segments, harmonics-to-noise ratio, position and bandwidth of 6 formants (resonance frequencies of the vocal tract depending strongly on its actual shape), 16 linear predictive coding coefficients, 12 mel frequency cepstrum coefficients ("spectrum of the spectrum"), and 12 linear frequency cepstrum coefficients (without the perceptually oriented transformation into the mel frequency scale), see Tab 1.

Tab 1. Basic acoustic feature contours (low level descriptors, LLD, based on frame-level)

Frame level based feature	Number of contours
Fundamental frequency	1
Intensity	1
Harmonics-to-noise ratio	1
Linear predictive coding (1-16)	16
Formants (F1-F6) position	6
Formants (F1-F6) bandwidth	6
Voiced segments duration	1
Unvoiced segments duration	1
Mel frequency cepstrum coefficients (MFCC, 1-12)	12
Linear frequency cepstrum coefficients (LFCC, 1-12)	12

After splitting the speech signal into frames and computing the above mentioned frame level features, the values of each frame level feature are connected to contours. This procedure results in 57 speech feature contours (e.g. the fundamental frequency contour, the bandwidth of formant 4 contour etc.), which are joined by their first and second derivates (velocity (Δ) and acceleration ($\Delta\Delta$) contours). Furthermore these 171 speech feature contours are described by elementary statistics (linear moments, values and positions of extrema, quartiles, ranges, length of time periods beyond threshold values, regression coefficients, etc.), and spectral descriptors (spectral energy of low frequency bands vs. high frequency bands, etc.) resulting in 8,550 high-level speech features (171 speech contours x 50 functionals), see Fig 1.

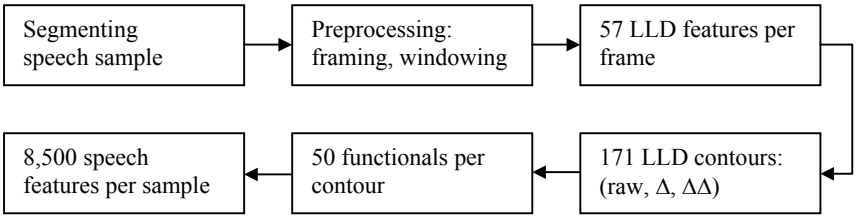

Fig. 1. Processing flow of acoustic feature computation including the computation of frame level based features and contour descriptors (functionals) to capture sufficient temporal information.

3 Speech Adapted Pattern Recognition Framework

The acoustic measurement process follows the speech adapted steps of pattern recognition: (a) recording speech, (b) preprocessing, (c) feature computation, (d) dimensionality reduction, (e) classification, and (f) evaluation. The following listing gives a brief overview about possible variations in the measurement process.

Recording speech: *Source of verbal material* [human-to-human, human-to-machine communication; monologue vs. dialogue situations; speech databases (e.g. AEC [16], Sympafly [17], EMO-DB [18])]; *Speaking format* [vowel phonation, isolated words, connected speech, read speech, spontaneous speech]; *Speaking style* [intensity and articulation related speaking style (e.g. hyperarticulation, whispering, shouting)]; *Speech segment* [vowels, consonant types (fricative, stop, glide), consonant clusters, syllables, words, chunks phrases]; *Recording situation* [noisy vs. noise subdued environment (e.g. driving with open window vs. laboratory recording); rough vs. clean speech signal quality (e.g. telephone call, radio communication vs. clean recording in 22.05 kHz, 16 bit)].

Preprocessing: *Segmentation* [manual, (semi-)automatic segmentation (e.g. MAUS system [19]) of the speech signal in phonetic units of interest (e.g. specific vowels, types of consonants or consonant cluster, stressed syllables, beginning or end of phrases)]; *Noise reduction* [outlier detection, moving average filter, low bandpass filter]; *Framing and Windowing* [size of frames (10-20 ms), degree of overlapping, window function (hamming, hanning)].

Feature computation. *Low level descriptors (LLD)* [Fundamental frequency, intensity, harmonics-to-noise ratio, formant position and bandwidth (F1-F6), LPC, MFCC, LFCC, partitioning into voiced and unvoiced speech segments]; *Functionals* [elementary statistics (e.g. linear moments, extrema values and positions, quartiles, ranges, length of time periods beyond threshold values, regression coefficients), spectral descriptors (e.g. spectral energy of low frequency bands vs. high frequency bands) and state space features (e.g. largest lyapunov coefficient)]; Automatic feature generation (genetic algorithms); *Normalization* [individual speaker specific baseline correction, age/ gender specific normalization].

Dimensionality reduction. *Subset selection* [supervised filter-based (e.g. correlation, information gain ratio), unsupervised (e.g. density, entropy, salience) or wrapper-based subset selection (forward selection, backward elimination, sequential forward floating search, genetic algorithm selection)]; *Feature Transformation* [unsupervised (e.g. Principle Component Analysis, PCA Network, Nonlinear Autoassociative Network, Multidimensional Scaling, Kernel PCA, Independent Component Analysis, Sammon Map, Enhanced Lipschitz Embedding, SOM, Spectral Transformation, Wavelet Transformation); supervised (e.g. Linear Discriminant Analysis)].

Classification. *Classification granularity* (binary or multiple class prediction)*; Classifier choice* [e.g. 1-nearest neighbour, multi-layer perceptron, support vector machine, linear discriminant analysis, hidden markov model, decision tree, gaussian

mixture model]; *Metaclassifier* [bagging, boosting, voting, stacking]; *Parameter optimization*.

Validation. *Evaluation strategy* [2-, 5-, 10-fold cross validation; leave-one-sample-out]; *Reliability strategy* [recordings on different days for retest reliability (e.g. leave-one-session-out)].

4 Fatigue Related Speech Changes

The following fatigue and sleepiness related physiological changes can influence voice characteristics: (a) decreased muscle tension (reduced facial expression and smiling, unconstricted pharynx, softening of vocal tract walls, vocal fold elasticity and tension), (b) decreased body temperature (reduced heat conduction, changed viscoelasticity of vocal folds, changed friction between vocal tract walls and air as well as impaired laminar flows), (c) reduced cognitive processing speed (impaired speech planning) and (d) flat and slow respiration (low subglottal pressure). The corresponding acoustical effects are lower fundamental frequency, intensity, articulatory precision, and rate of articulation, as well as shift in the spectral energy distribution due to changed filter characteristics. These filter characteristics can be described in terms of formant frequencies and bandwidths. Formant bandwidth is among others determined by the amount of energy loss in the vocal tract due to softening of vocal tract walls, raising of viscosity, and increasing heat conduction [20,21]. It is expected that especially the softening of vocal tract walls produces a moderate increase in formant frequencies and broadening of formant bandwidths, especially in lower formants. In an antagonistic way, the friction between air and the vocal tract walls (due to a lowered body temperature and lowered heat conduction) may cause a lowering of the formant frequencies.

Furthermore a general decreased facial expression and thus decreased lip spreading [22,23] may result in a shortening of the vocal tract and therefore in lower F1 and F2 frequencies. A supplemental factor for causing this effect can be seen in articulatory effects. The reduction of articulatorical effort leads to a smaller opening degree during slackened articulation and a decreasing of the first formant. Another sleepiness related voice phenomenon, which might influence the formant values, is the shift of speech quality to a breathy, wide, lax, and non-tensed voice. The tensing of the vocal tract raises the larynx, which could result in an increased formant frequency. On the contrary, a lax voice is characterised by low F1 and wide F1 bandwidth. Similarly, the breathier a vowel is spoken, the wider is the first formant bandwidth. However, little empirical research has been done to examine the effect of fatigue and sleepiness on acoustic voice characteristics. The aim of the following study is to introduce a fatigue detection method based on the speech adapted pattern recognition approach.

5 Empirical Validation Results

We conducted a within-subject sleep deprivation design (N = 21; 8.00 p.m to 4.00 a.m). During the night of sleep deprivation a well established, standardised self-report sleepiness measure, the Karolinska Sleepiness Scale (KSS) was used every hour just before the speech recordings. The verbal material consisted of a 2 second sustained phonation of the German vowel [a:]. The participants recorded other verbal material at the same session, but in this article we focus on sustained phonation only. For training and classification purposes, the records were further divided in two classes: sleepy (SS) and non sleepy (NSS) with the boundary value KSS ≥ 6. (4 samples per subject; 2 samples recorded at 8.30 and 9.00 p.m and 2 samples recorded at 3.00, 3.30 a.m.; total number of speech samples: 84 samples; 61 samples NSS, 23 samples SS). During the night, the subjects were confined to the laboratory and supervised throughout the whole period. Between sessions, they remained in a room, watched DVD, and talked. Non caffeinated beverages and snacks were available ad libitum.

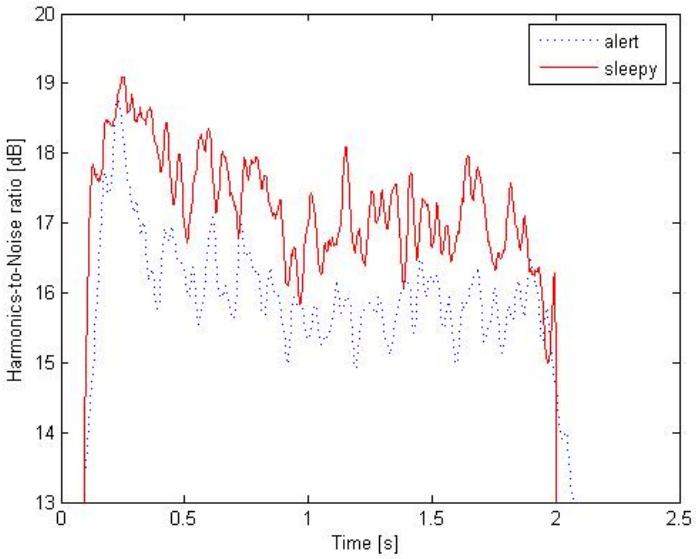

Fig. 2. Average harmonics-to-noise ratio contours of the sustained german vowel [a:] phonation for sleepy vs. alert speakers.

The recognition rate (number of cases classified correctly divided by all cases; RR) of the 1-nearest neighbour classifier was 83.3%, the class-wised averaged classification rate (mean of recognition rates for each class; CL) was 76.5%. The five best single features have the following correlations to self-reported sleepiness: HNR_mean = .29 (see Fig 2), formant4_max = -.29, formant2_bandwidth_1.quartil = .29, intensity_1.quartil = .28, and formant3_max = -.28.

6 Concluding Remarks

Due to the hypothesized sleepiness related physiological changes in cognitive speech planning, respiration, phonation, articulation, and radiation, the results for the reported classification performance above chance level were largely as could be expected. This is consistent with previous sleepiness related findings that suggest an association of acoustic features [24, 25] with sleepiness. Nevertheless our results are limited by the facts that we did not consider real life speaking situations including (a) variations in speaking format (e.g. read speech, spontaneous speech), (b) variation in speakers´ states (e.g. having a cold, after drinking milk, being nervous, aggressive or in a depressive mood), (c) variations in speakers´ trait (e.g. strong dialect, older age), and (d) variations in situational context factors (e.g. noisy environments, room microphone). These confounders might influence the detection rate and the false alarm error rate of the sleepiness measurement. Thus the present results are preliminary and need to be replicated using natural speech environment. Moreover, it would seem advisable that future studies address the following topics:

- different validation designs: following the circadian sleepiness cycle, pharmacological studies, randomized controlled trials (between-subject designs).
- temporal segmentation: finding sleepiness sensitive phonetic units (vowels or consonant cluster in different positions within words and phrasal units).
- feature extraction: computing state space domain based features (e.g. average angle or length of embedded space vectors, lyapunov exponents, correlation dimension, automutual information, time resolved density, fractal dimensions, multiscale entropies, and recurrence quantification analysis); using evolutionary feature generation.
- classification: utilizing maximum-likelihood bayes classifiers, fuzzy membership indexing, HMMs, gaussian mixture density models.

References

1. MacLean, A.W.: Sleepiness and Driving, Sleep Medicine Reviews *7*, (2003) 507-521.
2. Melamed, S.: Excessive Daytime Sleepiness and Risk of Occupational Injuries in Non-Shift Daytime Workers", Sleep 25(3), (2002) 315-322.
3. Wright, N., McGown, A.: Vigilance on the Civil Flight Deck: Incidence of Sleepiness and Sleep during Long-Haul Flights and Associated Changes in Physiological Parameters, Ergonomics 44, (2001) 82-106.
4. Durmer, J.S., Dinges, D.F.: Neurocognitive Consequences of Sleep Deprivation. Seminars in Neurology, 25, 2005. 117-129.
5. Nilsson, J.P., Soderstrom, M., Karlsson, A.U., Lekander, M., Akerstedt, T., Lindroth, N.E., Axelsson, J.: Less Effective Executive Functioning after one Night´s Sleep Deprivation. Journal of Sleep Research, 14, 2005. 1–6.
6. Cañas, J.J.: Technology for special needs. An Interdisciplinary Journal on Humans in ICT Environments, 2, 2006, 4–7.

7. Kollias, S., Amir, N., Kim, J., Grandjean, D.: Description of Potential Exemplars: Signals and Signs of Emotion. HUMAINE Human-Machine Interaction Network on Emotions. (2004).
8. Caffier, P.P.: The Spontaneous Eye-Blink as Sleepiness Indicator in Patients with Obstructive Sleep Apnoea Syndrome-a Pilot Study, Sleep Medicine, 2, (2002). 155-162.
9. Sommer, D., Chen, M., Golz, M., Trunschel, U., Mandic, D.: Fusion of State Space and Frequency Domain Features for Improved Microsleep Detection. In: W. Dutch et al. (Eds.): Int Conf Artifical Neural Networks (ICANN 2005), Springer: Berlin (2005) 753-759.
10. Vöhringer-Kuhnt, T., Baumgarten, T. Karrer, K., Briest, S.: Wierwille's Method of Driver Drowsiness Evaluation Revisited. Proceeding of International Conference on Traffic & Transport Psychology. (2004).
11. Schuller, B.: Automatische Emotionserkennung aus sprachlicher und manueller Interaktion. [Automatic Emotion Recognition from verbal and manual Interaction]. Dissertation, Technische Universität München. (2006).
12. Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. Proceedings of Interspeech, (2007) 2253-2256.
13. Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G.: Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech. Proceedings of Interspeech, (2007) 2249-2252.
14. Batliner, A., Steidl, S., Schuller, B, Seppi, D., Laskowski, K, Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: Combining Efforts for Improving Automatic Classification of Emotional User States. In: Erjavec, T. & Gros, J.Z. (Eds.): Language Technologies, IS-LTC 2006, Ljubljana, Slovenia: (2006) 240-245
15. Mierswa, I., Morik, K.: Automatic Feature Extraction for Classifying Audio Data. Kluwe, Amsterdam (2005).
16. Batliner, A., Hacker, C., Steidl, S., Noeth, E., D'Arcy, S, Rusell, M., Wong, M.: "You stupid tin box" – Children interacting with the AIBO robot: A crosslinguisitc emotional speech corpus, Proceedings of the 4th International Conference of Language Resources and Evaluation LREC 2004 (LREC Lisbon 2004) (2004) 171-174.
17. Steidl, S., Hacker, C., Ruff, C., Batliner, A., Noeth, E., Haas, J.: Looking at the Last Two Turns, I'd Say This Dialogue is Doomed – Measuring Dialogue Success ,Proceedings TSD (Text, Speech and Dialog) (2004) 629-636.
18. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech, Proceedings of Interspeech 2005 (Lisboa, Portugal) (2005) 1517-1520.
19. Schiel, F.: MAUS Goes Iterative. Proc. of the IV. International Conference on Language Resources and Evaluation, Lisbon, Portugal, (2004) 1015-1018.
20. Rabiner, L., Schafer, R.W.: Digital Processing of Speech Signals (Prentice-Hall, Upper Saddle River, New Jersey, USA (1978).
21. Scherer, K.R.: Vocal affect expression: A review and a model for future research, Psychological Bulletin, 99 (1986) 143-165.
22. Kienast, M., Sendlmeier, W.F.: Acoustical analysis of spectral and temporal changes in emotional speech, Speech Emotion (2000) 92-97.
23. Tartter, V.C.: Happy talk - Perceptual and acoustic effects of smiling on speech, Perception and Psychophysics, 27(1) (1980) 24-27.
24. Nwe, T.L., Li, H**.,** Dong, M.: Analysis and Detection of Speech under Sleep Deprivation. Proceeding of Interspeech, (2006) 17-21.
25. Krajewski, J., Kröger, B.: Using prosodic and spectral characteristics for sleepiness detection. Interspeech Proceedings, (2007) 1841-1844.