CLASSIFICATION OF SPEECH UNDER STRESS AND COGNITIVE LOAD IN USAR OPERATIONS

Marcela Charfuelan, Geert-Jan Kruijff

DFKI GmbH, Language Technology Lab, Berlin and Saarbruecken, Germany.
marcela.charfuelan@dfki.de, gj@dfki.de

ABSTRACT

This paper presents the classification of speech under stress and cognitive load in speech recordings of Urban Search and Rescue (USAR) training operations. The type of stress encountered in the USAR domain, more specifically in the human team communication, includes both physical or psychological stress and cognitive load. We were able to annotate and identify these two types of stress in recordings of real USAR training operations. Different acoustic features are extracted at full and subband level, SVM and adaptive GMMs are used as classifiers. Two strategies to improve the classification of speech under stress, in particular physical stress, are proposed. We have achieved a classification accuracy of 74% for three very unbalanced classes (physical stress, cognitive load and neutral), with 82% classification of physical stress.

Index Terms— stress, cognitive load, feature extraction, subband processing, speech classification.

1. INTRODUCTION

Rescue work in Urban Search and Rescue (USAR) operations is a physiologically, cognitively, and psychologically demanding task [1, 2]. Although USAR team members are skilled and highly trained people, they might be affected by stressors like: difficult perception (due to darkness, smoke or dust), lack of reliable communications, time pressure, cognitive fatigue, or emotional demands [1]. Thus, the type of stress encountered in the USAR domain, more specifically in the human team communication, includes both cognitive load and physical or psychological stress. Cognitive task load and affective task load have been identified in [2] as important factors to recognise critical states in geo-collaboration teamwork. In the USAR domain, identification of levels of cognitive load and affective task load in the form of physical stress, can be used to improve the mechanisms for scheduling and allocation of tasks. Although cognitive load and stress are two concepts that describe similar process [3], these two types of stress can be measured using physiological sensors [4, 5], but also through speech analysis [6, 7]. In the USAR domain, where complex rescue operations take place, it would be advantageous to use unobtrusive methods like speech analysis to detect stress.

Several studies have reported the analysis and classification of various levels of cognitive load and stress in different situations using speech analysis. In [8] a monitor system that evaluates indicators of fatigue and excessive demand on speech of both traffic controllers and pilots is proposed; in [9] the effect of three different types of cognitive load on speech prosody of pilots performing a task in a flight simulator is analysed; in [10] the speech produced by subjects while driving in a simulator at variable speed and engaged on mental tasks of variable cognitive load, is classified according to various stress categories. Most of the works on cognitive load and/or stress classification, concentrate the analysis in only one type of stress and use data recorded in controlled experiments or simulations. Very few works report analysis of real data, for example: in [11] a corpus extracted from the cockpit voice recorder of a crashed aircraft is studied; in [12] recordings of drivers' speech under cognitive load and frustration are analysed; and in [13] very noisy helicopter cockpit speech recordings are analysed.

In this paper we classify both physical stress and cognitive load in noisy speech data collected during USAR training sessions. We propose a two steps classification approach, where appropriate features are selected to train two classifiers working in tandem, one for classifying physical stress and another for classifying cognitive load. We show that a two step classification approach, gives better results than only one classifier trained with thousands of features. In the following, first we review robust acoustic features to classify either stress or cognitive load. Then we describe the data and stress annotations (Section 3) and the full band and subband acoustic features used in this study (Section 4). Classification experiments are presented in Section 5 and conclusions and future work in Section 6.

2. ROBUST ACOUSTIC FEATURES FOR STRESS AND COGNITIVE LOAD CLASSIFICATION

One approach that has been shown to be robust for analysing speech under stress in real situations is the multi-band processing of speech. Hansen et al. [7] have developed an acoustic measure based on multi-band non-linear processing of speech: the autocorrelation envelope of the critical band filtered Teager Energy Operator (TEO-CB-AutoEnv). This measure has been used to recognise simulated and actual speech under stress from the SUSAS database [14]. Cepstral coefficients extracted at subband level in [15], are also shown to be robust for classifying several levels of cognitive load in the presence of noise. Combination of features has also been investigated: in [8] TEO-CB-AutoEnv, prosody, voice quality and spectral features are combined to monitor stress in speech from air traffic controllers in a simulation experiment. Mel frequency cepstrum coefficients (MFCCs) and TEO-CB-AutoEnv features are also combined in [16], in a fusion scheme that improves physical stress classification in a controlled experiment.

In a previous work [17], we have analysed the acoustic correlates of two levels of annotated stress in the USAR training recordings. The results of this preliminary study indicate that mostly prosody and TEO features correlate with the high stress annotated level and characterise physical stress; whereas spectral, TEO and prosody correlate with the medium stress annotated level and characterise cog-

The work reported in this paper has received funding from the EU-FP7 ICT 247870 NIFTi project.

nitive load. In order to classify speech under stress and cognitive load in the USAR training recordings we propose the combination of TEO-AutoEnv, voice strengths and spectral features extracted at subband level; we also use classic prosodic and articulatory features, spectral features and voice quality features, recently reported as good discriminators of cognitive load [18]. For comparison we train a classifier with state of the art acoustic features (thousands of features used in the first Paralinguistic Challenge [19]) and also train an adaptive Gaussian Mixture Models (GMMs) classifier with MFCCs and their first and second derivative [20]. According to our results and the results reported in [12] it seems that careful selection and combination of a reduced number of appropriate features result in better classification performance, in particular when working with sparse and unbalance data.

3. DATA COLLECTION AND ANNOTATION

The speech database analysed in this paper corresponds to the recordings of the NIFTi Join Exercises 2011 on human-robot-teaming (NJEx2011) [21]. The NIFTi Join exercises took place in a constructed, complex environment where four different teams performed several missions in two days. On the first day (0706) each team had two missions: in mission 1 the teams traversed a complex arena with an unmanned ground vehicle (UGV), helped by an unmanned aerial vehicle (UAV); each team got 45 minutes. In mission 2 the teams explored two floors in the Red Building searching for victims; each team got 75 minutes. On the second day of exercises (0707) the teams went into the Red Building again but this time under more severe circumstances: smoke, fire, more floors to explore and in less time. Each team explored three floors of the Red Building searching for victims; each team got 90 minutes. In all the exercises UGV operation was remote, UAV was Line Of Sight (LOS) and the communication was done in English via open voice loop only. 7 sessions (missions) were recorded during the first day and 4 during the second day. Different team players (only male) participate in each session. The recordings of each session were segmented per turn and annotated according to the speakers, or team players, that participate on the mission. Table 1 shows the distribution of turns (utterances) per day and speaker

	Day	
Speaker	0706	0707
missionDirector	161	272
safetyDirector	817	324
teamRole	47	25
uavPilot	31	48
ugvPilot	343	197
whiteCommand	53	36
Total time	410 min.	315 min.

Table 1. NJEx2011 distribution of turns per day and speaker. The average duration of a turn is 6 seconds.

The segmented sessions were further annotated according to three levels of stress: level (1) no stress, speech is neutral, normal, relax, happy; level (2) medium stress, speech is nervous, there is tension in the voice, more speed, there are hesitations; and level (3) high stress, there are shouts, anger, despair. Three people annotated each utterance according to these levels. Full agreement of annotators was obtained in 69.6% of the data, additionally majority agreement (more than two annotators agree) was obtained on 29.5% of the data. In a previous experiment where the data was annotated by two people [17], we have selected the full agreement data for analysis, in this case the full agreement data set was smaller, in particular for the sparse classes: physical stress and cognitive load. So we decided to use an approach similar to the sparse-instances-based active learning reported in [22] to select more samples from the majority agreement set. In our case we start creating a classifier with the full agreement data, select K sparse samples from the majority agreement set, add these samples together with the majority agreement label to the training data and update the classifier. We repeat this procedure until there are no more sparse samples in the majority agreement set. Following this procedure the classifier is improved every time sparse samples are added. Once there are no more sparse samples to add we use the final classifier to select samples where the majority agree with the classifier prediction. In this way we discard outlier samples. The final distribution of data according to speakers and three stress categories is presented in Table 2.

Speaker	Physical stress (high stress)	Cognitive load (medium stress)	Neutral
missionDirector	0	16	401
safetyDirector	27	189	855
teamRole	0	4	67
uavPilot	0	1	77
ugvPilot	1	19	495
whiteCommand	0	2	85
Total	28	231	1980
Percentage	1.2%	10.3%	88.4%

Table 2. NJEx2011 distribution of turns per speaker type and annotated stress level. In a previous study [17], the two levels of annotated stress were identified as physical stress and cognitive load.

4. ACOUSTIC FEATURES

Three sets of acoustic features were extracted from the data and used on different classification experiments: (1) The first set includes 12 MFCCs and their first and second derivative, extracted at frame level using HTK [23]. MFCCs have been reported as good correlates of stress and cognitive load in recent literature [16, 12]. (2) In a second set we use state of the art features used in the classification of paralinguistic features [19]. In this case we use the openSMILE tool and the emobase2010 scripts to extract 1583 features that include: pitch, loudness, MFCC, log Mel Frequency Bands (MFB), Line Spectral Frequencies (LSP), voice quality features (jitter and shimmer) and functionals [24]. (3) In the third set of features, we extract full band and subband features reported as robust for analysing speech under stress in real situations. These features were mainly extracted using the snack toolkit [25]. In the following we briefly describe this third set of features:

4.1. Full band acoustic features

Among the full band features we have extracted standard prosodic features, spectral, articulatory and voice quality features:

- Standard prosodic features (extracted frame based): Fundamental frequency or pitch (F0); maximum, minimum, and range of F0; duration of the utterance in seconds; voicing rate calculated as the number of voiced frames (frames for which F0 > 0) per time unit; and log power calculated as the logarithm of the averaged short term energy: log_pow$=$ $\log(\frac{1}{N}\sum x^2)$, N is the length of the window frame.

- Spectral and articulatory-based features
  - Mel-cepstral coefficients
  - Formants: F_1, F_2, F_3, F_4
  - Formant bandwidths: B_1, B_2, B_3, B_4
  - Formant dispersion: calculated as: $FD = \frac{(F2-F1)+(F3-F2)+(F4-F3)}{3}$
- Voice quality (VQ) features: the following gradient VQ features, extracted at frame level, are rough spectral estimates of traditional voice quality parameters normally calculated in the time domain. These features were developed in [26] and were shown to be robust on the classification of emotions under different levels of noise and reverberation:
  - Open Quotient Gradient $= (\tilde{H}_1 - \tilde{H}_2)/F0$
  - Glottal Opening Gradient $= (\tilde{H}_1 - \tilde{A}_{1p})/(F_{1p} - F0)$
  - Skewness Gradient $= (\tilde{H}_1 - \tilde{A}_{2p})/(F_{2p} - F0)$
  - Rate of Closure Gradient $= (\tilde{H}_1 - \tilde{A}_{3p})/(F_{3p} - F0)$
  - Incompleteness of Closure $= B_1/F_1$

  These measures are gradients instead of amplitude ratios; they are calculated on the basis of frame-based raw measures like formant frequencies, formant bandwidths, amplitude of the first two harmonics at F0 and $2F0$: H_1, H_2, frequency of spectrum peaks near formants: F_{1p}, F_{2p}, F_{3p}, and amplitude of spectrum peaks near formants: A_{1p}, A_{2p}, A_{3p}. A tilde on some of the raw measures indicates vocal tract influence compensation. We also extract VQ features at utterance level based on the long term average spectrum (LTAS) in three bands of frequency [27]: 0-2kHz, 2-5kHz and 5-8kHz. For each of these bands the maximum LTAS level is selected.
  - Hamm_effort = LTAS_{2-5k}
  - Hamm_breathy = (LTAS_{0-2k} - LTAS_{2-5k}) - (LTAS_{2-5k} - LTAS_{5-8k})
  - Hamm_head = (LTAS_{0-2k} - LTAS_{5-8k})
  - Hamm_coarse = (LTAS_{0-2k} - LTAS_{2-5k})
  - Hamm_unstable = (LTAS_{2-5k} - LTAS_{5-8k})
  - slope_LTAS: least squared line fit of LTAS in the log-frequency domain (dB/oct)
  - slope_LTAS1kz: least squared line fit of LTAS above 1 kHz in the log-frequency domain (dB/oct)
  - slope_spectrum1kz: least squared line fit of spectrum above 1 kHz (dB/oct).

4.2. Subband acoustic features

The following subband features were implemented using the snack toolkit library [25]. First the speech signal is filtered with five band-pass filters with pass-bands: 0-1kHz, 1-2kHz, 2-4kHz, 4-6kHz and 6-8kHz. Then the following features are calculated from each band-pass signal:

- Teager Energy Operator - autocorrelation envelope (TEO-AutoEnv) [28]: this is a measure that has been used to detect and classify speech under stress (emotional, task stress, Lombard effect) in the SUSAS database. The Teager operator for a discrete-time signal is defined as [28]:
$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1)$$
In our implementation of the TEO-AutoEnv, we apply the TEO operator to the five filtered signals, then the autocorrelation from each TEO band is calculated and the area under the autocorrelation envelope is calculated and normalised over the window lag. The TEO operator has been applied to the five filtered signals, instead of the 16 critical bands proposed in [28], because it has been found that some bands are less sensitive for stress/neutral speech classification [29], so we reduce the number of bands and extract other features on each band to study their redundant or complementary effect.
- Voicing strengths (STR): estimated with peak normalised cross correlation of the input signal. The correlation coefficient for a delay t is defined by:
$$c_t = \frac{\sum_{n=0}^{N-1} s(n)s(n+1)}{\sqrt{\sum_{n=0}^{N-1} s^2(n) \sum_{n=0}^{N-1} s^2(n+t)}}$$
In a previous work [30], we have found that voicing strengths are correlated with vocal effort of dominant speech, so it is expected that these features are correlated as well with some type of stressed speech (shouting, angry speech, etc.).
- Spectral entropy (SPE): this feature is a kind of "peakiness" of the spectrum. This feature is calculated as follows [31]: the spectrum X is converted into a Probability Mass Function (PMF) normalising it by: $x_i = \frac{X_i}{\sum_{i=1}^{N} X_i} \quad i = 1 : N$ where X_i is the energy of the i^{th} frequency component of the spectrum, x is the PMF of the spectrum and N is the number of points in the spectrum. Entropy for each frame is calculated by: $H(x) = -\sum_{x \in X} x_i * log_2 x_i$
Spectral entropy has been used in speech endpoint detection and in classification of emotions.

5. CLASSIFICATION

In order to tackle the problem of classification of very unbalanced data, several strategies have been proposed. In this work we have used weighted support vector machine (SVM), where the weight values are determined by the proportion of data in each class. Two weighted SVM classifiers were trained and tested, one was trained with 1583 openSMILE features and the other with the full-band and subband features described in Section 4. Another strategy to classify very unbalanced data, is to create a universal or background model, using for example the neutral data, and adapt it to the acoustics characteristics of the sparse sets [20]. In this work, we have created a background Gaussian Mixture Model (GMM) using the annotated neutral data. This is a GMM of 32 mixtures trained with MFCCs features and their first and second derivatives. Maximum likelihood linear transformations (MLLR) and Maximum a-posteriori (MAP) adaptation techniques [23] were used to adapt the background model with physical stress and cognitive load data. Classification is performed using the two adapted models and the background neutral model. Two classification experiments were designed as follows:

1. **Speaker dependent classification (SD):** 40 repetitions of stratified sampling where 2/3 of the data in each class is randomly selected to train the models and the other 1/3 (not used for training) is used for testing. Stratified sampling is used in order to keep a balance on the amount of data for training and testing each class.
2. **Speaker independent classification (SI):** in this case the safetyDirector speaker of each session, who is a different person in every session, is used for testing and the rest of the

data for training. Classification results of all safetyDirector speakers of all sessions are averaged for a final SI score.

Baseline classification results for physical stress, cognitive load and neutral speech data are presented in Table 3. Best classification results in terms of a good compromise between individual classes classification and overall accuracy are presented in bold. Similar results for physical stress and neutral are obtained with the **FULL-SUBBAND** and **MFCC_0_D_A** classifiers. The **EMOBASE2010** classifier gives good classification for cognitive load and neutral but almost random for physical stress, which might be due to the small amount of physical stress samples to train a classifier with so many features.

Classifier		Features	PS	CL	N	Acc.
SD	**SVM-r**	**FULL-SUBBAND**	**45.8**	**76.9**	**81.3**	**80.4**
	32GMMs	MFCC_0_D_A	45.8	70.1	74.2	73.4
	SVM-p	EMOBASE2010	31.1	83.4	73.5	74.0
SI	**SVM-r**	**FULL-SUBBAND**	**48.1**	**52.4**	**74.9**	**70.2**
	32GMMs	MFCC_0_D_A	55.6	45.0	64.3	60.7
	SVM-p	EMOBASE2010	25.9	77.2	73.6	73.0

Table 3. Baseline classification of physical stress (PS), cognitive load (CL) and neutral (N) data. Acc: overall classification accuracy, SD: speaker dependent, SI: speaker independent. SVM-r: radial kernel, degree 3; SVM-p: polynomial kernel, degree 1.

5.1. Two steps classification approach

In this experiment motivated by the analysis presented in [17], where it was found that the acoustic correlates of physical stress and cognitive load are very different, we decided to perform the classification of the three classes in two steps. In the first step we classify physical stress and the rest of the data (both cognitive load and neutral) using just subband features. Subband features were found to be the best discriminant features between physical stress and the rest of the data. In a second step we classify cognitive load and neutral data. We have found that in our data the features that better discriminate speech under cognitive load and neutral speech are VQ and MFCCs. We compare results using MFCC_0_D_A and EMOBASE2010 features in the second step. The results of the two steps classification approach are presented in Table 4. These results include error correction due to misclassifcations in the first step.

Classifier/features						
	STEP 1	STEP 2	PS	CL	N	Acc.
SD	SVM-r / SUBBAND	**SVM-r/VQ-MFCC**	**75.6**	**54.3**	**77.5**	**75.1**
		32GMMS/MFCC_0_D_A	78.0	48.6	73.3	70.8
		SVM-p/EMOBASE2010	73.6	55.7	68.9	67.6
SI	SVM-r / SUBBAND	**SVM-r/VQ-MFCC**	**70.4**	**50.3**	**67.7**	**64.7**
		32GMMs/MFCC_0_D_A	70.3	43.9	61.4	58.5
		SVM-p/EMOBASE2010	70.4	38.1	62.2	58.2

Table 4. Two steps classification approach of physical stress (PS), cognitive load (CL) and neutral (N) data. Acc: overall classification accuracy, SD: speaker dependent, SI: speaker independent. SVM-r: radial kernel, degree 3; SVM-p: polynomial kernel, degree 1.

5.2. Adding controlled data to boost stress classification

As can be seen in Table 3 the classification of physical stress data is particularly low, which might be due to the very little amount of data in this class. The characteristics of this type of data are clearly distinguishable, that is, typical shouting, anger or despair of physical stress. This is a type of stress that can be found in the SUSAS database, which in our case fit very well because there is available male samples and in English. So, in order to boost the classification of physical stress we have added some samples of simulated anger from the SUSAS database to the training data. We have selected utterances longer than one second, from four speakers: 17 samples of anger and 18 samples of neutral speech data. Adding these samples to our best classifier in Table 4 improved the classification of physical stress and cognitive load as shown in Table 5.

Classifier/features						
STEP 1	STEP 2		PS	CL	N	Acc.
SVM-r / SUBBAND	SVM-r / VQ-MFCC	SD	82.7	60.2	75.6	74.2
		SI	66.7	50.1	67.1	65.2

Table 5. Classification of physical stress (PS), cognitive load (CL) and neutral (N) data using two steps and adding simulated anger and neutral samples from the SUSAS database. SD: speaker dependent, SI: speaker independent. SVM-r: radial kernel, degree 3.

6. CONCLUSIONS

We have presented the classification of speech under stress and cognitive load in speech recordings of USAR training operations. In contrast to most of the analysis of speech under stress and/or cognitive load reported in the literature, we have analysed speech recordings of real situations under very noisy conditions. The stress levels in this data were determined by manual annotation and not by the recording condition or experimental setting. Speaker dependent and speaker independent classification experiments were performed. A speaker dependent solution might be meaningful in cases where data from the rescue team is available, which might be the case when designing a system for a particular team.

We have proposed to handle the speech signal at subband level as follows: the well known, stress correlated, TEO-AutoEnv feature is extracted at subband level and combined with voice strengths and spectral features, also extracted at subband level. The features extracted at subband level proved to be more robust when compared to the full-band features or the thousands of features extracted in a "brute-force" approach. In particular the proposed subband features were found to be robust to classify physical stress.

Two strategies to improve the classification of stress were proposed: a two steps classification approach and the adding of controlled data from the SUSAS database to boost stress classification. We have shown that the two steps classification approach, where appropriate features are selected to train two classifiers working in tandem, gives better results than only one classifier trained with hundreds of features. We have compared our results with state of art methods and techniques like adaptive GMMs and SVM classifiers trained with thousands of features. We have achieved a speaker dependent classification accuracy of 74% for three very unbalanced classes (physical stress, cognitive load and neutral), with 82% classification of physical stress.

In future work we will consider to use speaker normalisation in order to improve speaker independent classification. Also in order to improve cognitive load classification (the second step in our two steps classification approach), we will consider combining acoustic and linguistic features like hesitations, fluency problems, silent pauses, filled pauses, which are expected to occur during time pressure and cognitive load [32].

7. REFERENCES

[1] R. R. Murphy, "Human-robot interaction in rescue robotics," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 2, pp. 138–153, 2004.

[2] R. Looije, G. te Brake, and M.A. Neerincx, "Geo-collaboration under stress," in *Workshop on Mobile HCI for Emergencies*, Singapore, 2007.

[3] A. W. K. Gaillard, "Comparing the concepts of mental load and stress," *Ergonomics*, vol. 36, no. 9, pp. 991–1005, 1993.

[4] Jennifer A. Healey and Rosalind W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, pp. 156–166, 2005.

[5] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psychophysiological measures for assessing cognitive load," in *12th ACM international conference on Ubiquitous computing*, Copenhagen, Denmark, 2010, Ubicomp '10.

[6] K. R. Scherer, D. Grandjean, T. Johnstone, G. Klasmeyer, and T. Bänziger, "Acoustic correlates of task load and stress," in *ICSLP2002 - Interspeech 2002*, Denver, USA, 2002.

[7] J. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker Classification I*, vol. 4343 of *Lecture Notes in Computer Science*, pp. 108–137. Springer Berlin / Heidelberg, 2007.

[8] J. Luig and A. Sontacchi, "Workload monitoring through speech analysis: Towards a system for air traffic control," in *27th Congress of The International Council of Aeronautical Sciences (ICAS)*, Nice, France, 2010.

[9] K. Huttunen, H. Keränen, E. Väyrynen, R. Pääkkönen, and T. Leino, "Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights," *Applied Ergonomics*, vol. 42, no. 2, pp. 348 – 357, 2011.

[10] R. Fernandez and R. W. Picard, "Modeling drivers' speech under stress," *Speech Commun.*, vol. 40, pp. 145–159, 2003.

[11] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, and D. Poch, "Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions," *Speech Communication*, vol. 20, pp. 111–129, 1996.

[12] H. Boril, S. O. Sadjadi, T. Kleinschmidt, and J. H. L. Hansen, "Analysis and detection of cognitive load and frustration in drivers' speech," in *Interspeech 2010*, Makuhari, Japan, 2010.

[13] B. Schuller, M. Wimmer, D. Arsic, T. Moosmayr, and G. Rigoll, "Detection of security related affect and behaviour in passenger transport," in *Interpeech*, Brisbane, Australia, 2008.

[14] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: a speech under simulated and actual stress database," in *Eurospeech*, Rhodes, Greece, 1997.

[15] P. N. Le, J. Epps, E. Ambikairajah, and V. Sethu, "Robust speech-based cognitive load classification using a multi-band approach," in *APSIPA Annual Summit and Conference (APSIPA)*, Biopolis, Singapore, 2010.

[16] S. A. Patil and J. H. L. Hansen, "Detection of speech under physical stress: Model development, sensor selection, and feature fusion"," in *Proceedings of Interspeech*, Brisbane, Australia, 2008.

[17] M. Charfuelan and G. J. Kruijff, "Analysis of speech under stress and cognitive load in USAR operations," in *Procedings of IWSDS 2012: Towards a Natural Interaction with Robots, Knowbots and Smartphones*, Paris, France, 2012.

[18] T. F. Yap, J. Epps, E. Ambikairajah, and E. Choi, "Voice source features for cognitive load classification," in *Proceedings of ICASSP*, Prague, Czech Republic, 2011.

[19] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language – state-of-the-art and the challenge," *Computer Speech And Language*, 2012.

[20] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," in *Proceedings of ICASSP*, Las Vegas, NV, USA, 2008.

[21] G.J.M. Kruijff, "Proceedings of NJEx 2011, NID 2011," unpublished report, February 2012.

[22] Z. Zhang and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Proceedings of Interspeech*, Portland, OR, USA, 2012.

[23] S. Young et. al., "The HTK Book (for HTK Version 3.4)," http://htk.eng.cam.ac.uk/prot-docs/htk_book.shtml, 2012.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. 2010, MM '10, pp. 1459–1462, ACM.

[25] K. Sjölander et. al., "The Snack Sound Toolkit," http://www.speech.kth.se/snack, 2012.

[26] M. Lugger, B. Yang, and W. Wokurek, "Robust estimation of voice quality parameters under realworld disturbances," in *Proceedings of ICASSP*, Toulouse, France, 2006.

[27] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice quality," *Acta Otolaryngologica*, , no. 90, 1980.

[28] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[29] J. H. L. Hansen, W. Kim, M. Rahurkar, E. Ruzanski, and J. Meyerhoff, "Robust emotional stressed speech detection using weighted frequency subbands," *EURASIP Journal on Advances in Signal Processing*, p. 10, 2011.

[30] M. Charfuelan and M. Schröder, "The vocal effort of dominance in scenario meetings," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[31] H. Misra, S. Ikbal, S. Sivadas, and H. Bourlard, "Multi-resolution spectral entropy feature for robust ASR," in *Proceedings of ICASSP*, Philadelphia, PA, USA, 2005.

[32] A. Jameson, J. Kiefer, C. Müller, B. Gromann-Hutter, F. Wittig, and R. Rummer, "Assessment of a user's time pressure and cognitive load on the basis of features of speech," in *Resource-Adaptive Cognitive Processes*. Springer Berlin Heidelberg, 2010.