Analysis of speech under stress and cognitive load in USAR operations

Marcela Charfuelan and Geert-Jan Kruijff

Abstract This paper presents on-going work on analysis of speech under stress and cognitive load in speech recordings of Urban Search and Rescue (USAR) training operations. During the training operations several team members communicate with other members on the field and members on the control command using only one radio channel. The type of stress encountered in the USAR domain, more specifically on the human team communication, includes both physical or psychological stress and cognitive task load. Physical stress due to the real situation and cognitive task load due to tele-operation of robots and equipment. We were able to annotate and identify the acoustic correlates of these two types of stress on the recordings. Traditional prosody features and acoustic features extracted at sub-band level probed to be robust to discriminate among the different types of stress and neutral data.

1 Introduction

For several years the applications of stress detection in speech were mainly related to improve speech recognition, speaker recognition, or to improve the naturalness of synthetic speech [3]. Nowadays applications including detection of speech under stress and/or cognitive load span many fields. In human-computer interaction (HCI) and human-machine interaction (HMI) there is an increasing interest in analysing stress in speech. For example, [5] explored the prospects of exploiting the user's speech as a source of evidence for the recognition of resource limitation. Models of cognitive task load (CTL) as well as models of affective task load (ATL) and performance level are proposed in [7] to recognise critical states, with the objective of enhancing geo-collaboration on teamwork. The type of stress encountered

Marcela Charfuelan
DFKI GmbH, Language Technology Lab, Berlin, Germany. e-mail: marcela.charfuelan@dfki.de

Geert-Jan Kruijff
DFKI GmbH, Language Technology Lab, Saarbrücken, Germany. e-mail: gj@dfki.de

in the Urban Search and Rescue (USAR) domain, more specifically on the human team communication, includes both physical or psychological stress and cognitive task load. Physical stress due to the real situation and cognitive task load due to tele-operation of robots and equipment. The expectation is that collaborative team work will benefit from the automatic detection of critical affective states (stress). For example in an application involving multiple sources of information, the control command might decide to adapt or limit the information presented to the team members when different stress conditions are detected.

One approach that has been shown to be robust to analyse speech under stress in real situations is the multi-band processing of speech. Hansen et al. [3] have developed an acoustic feature based on multi-band non-linear processing of speech: the autocorrelation envelope of the critical band filtered Teager Energy Operator (TEO-CB-AutoEnv). This feature has been used to recognise simulated and actual speech under stress from the SUSAS database [4]. In our study we have used traditional prosody features extracted at full band level, and TEO-AutoEnv, spectral and voicing strength features extracted at sub-band level.

The paper is organised as follows: first we briefly describe our experience collecting and annotating speech data from USAR training sessions (Section 2). Then we briefly describe the acoustic features (Section 3), how we use them to identify acoustic correlates of the annotated stress and preliminary stress classification results (Section 4). Conclusions and future work are presented in Section 5.

2 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) [6]. 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 on 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 via open voice loop only. 7 sessions (missions) were recorded during the first day and 4 during the second day. Different team players (persons) 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. The segmented sessions were further annotated according to three levels of stress: (1) unstress: normal or neutral speech, happy, relax; (2) stress: speech is nervous, there is tension in the voice, more

Speaker	Day	
	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.

Speaker	Higher	Medium	Neutral
missionDirector	0	13	375
safetyDirector	24	188	629
teamRole	0	4	63
uavPilot	0	1	74
ugvPilot	0	16	437
whiteCommand	0	4	79
Total	24	226	1657
Percentage	1.2%	11.8%	86.8%

Table 2 NJEx2011 distribution of turns per speaker type and annotated stress level, where the annotators agree.

speed, there are hesitations; and (3) very stressed: there are shouts, anger, despair. Two people annotated these three levels of stress on each utterance of all sessions. The distribution of data according to speakers and three stress categories: higher (stress level 3), medium (stress level 2) and neutral (stress level 1), is presented in Table 2. According to this table there is very small number of higher and medium stress turns, and in particular higher stress is only exhibited by the safetyDirector speaker of the sessions. The inter-rater agreement is presented in Table 3. The number of observed agreements is 1908 (81.02% of the observations) and the number of agreements expected by chance is 1553.1 (65.95% of the observations). The Kappa value is 0.443 with 95% confidence interval: from 0.401 to 0.484. The strength of agreement is considered to be “moderate”, although as reported by [1], kappa values between 0.4 and 0.7 are usually regarded as fair agreement in annotations of this type of expressive speech data. For the analysis of stress in this data we have selected the turns where the two annotators agree.

Stress level	Neutral	Medium	Higher	Total turns
Neutral	**1658**	287	2	1947
Medium	118	**226**	14	358
Higher	3	23	**24**	50
Total turns	1779	536	40	2355

Table 3 NJEx2011 stress annotation: two annotators inter-rater agreement, Kappa=0.443

3 Acoustic features

Standard prosodic features and TEO sub-band features reported in the literature as good correlates of stress were extracted from the data; these and other sub-band features were extracted with snack [11] and are described below:
(a) Standard prosodic features: 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= $\log(\frac{1}{N}\sum s^2)$ where N is the length of the window frame. Prosodic features are extracted frame based and at full-band.

(b) Teager Energy Operator - Autocorrelation Envelope (TEO-AutoEnv): this is a measure that has been used to detect and classify speech under stress (emotional stress, task load stress and Lombard effect) in the SUSAS database. The Teager operator for a discrete-time signal s is defined as [12]:

$$\Psi[s(n)] = s^2(n) - s(n+1)s(n-1)$$

Similarly to the TEO-AutoEnv measure proposed in [12], we have implemented five bandpass filters with pass-bands: 0-1kHz, 1kHz-2kHz, 2kHz-4kHz, 4kHz-6kHz and 6kHz-8kHz. 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.

(c) Voicing strengths (STR): estimated with peak normalised cross correlation of the input signal. The correlation coefficient for a signal s and 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 [2], 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.).

(d) Spectral entropy (SPE): is a kind of "peakiness" of the spectrum that has been used in speech endpoint detection and in classification of emotions. This feature is calculated as follows [8]: 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:

$$H(x) = -\sum_{x \in X} x_i * log_2 x_i$$

4 Acoustic correlates of higher and medium stress types

One of the objectives in this work is to get a better understanding of the acoustic characteristics of the annotated data. Analysis of variance (AOV) of the acoustic features described in Section 3, was performed in order to establish the main acoustic correlates of the two types of annotated stress. The idea is to find out which features are significantly different among the sets of data: higher (H), medium (M) and neutral (N). Results are present in Table 4. We have analysed which features are significantly different among the three classes (H/M/N), between the medium level and neutral (M/N) and between the higher level of stress and the medium and neutral levels together (H/(M&N)). We can observe in this table that most of the features, except voicing strengths in some bands are significantly different among the three classes (H/M/N) . Prosody, TEO-AutoEnv and spectral features in higher bands are significantly different between medium and neutral data (M/N). In average f0 for medium stress is greater than f0 for neutral speech; log_pow is also in average

Acoustic features			Stress types and Neutral		
			H / M / N	M / N	H / (M & N)
Full-band	(a) Prosody	f0	***	***	***
		max_f0	**	**	–
		min_f0	***	*	***
		range_f0	•	*	–
		dur_seconds	***	***	**
		voicing_rate	•	*	–
		log_pow	***	***	*
Sub-band	(b) Voicing strengths	str1	**	–	***
		str2	*	–	*
		str3	–	–	–
		str4	–	–	–
		str5	•	*	–
	(c) TEO-AutoEnv	teo1	–	–	–
		teo2	***	***	–
		teo3	***	***	***
		teo4	***	***	***
		teo5	***	***	***
	(d) Spectral entropy	se1	***	–	***
		se2	***	***	–
		se3	***	***	•
		se4	**	**	–
		se5	***	***	*
SVM classification accuracy (avg)			75%	76%	83%
Classification per class %			**H:**43 **M:**66 **N:**76	**M:**75 **N:**76	**H:**71 **(M&N):**83

Table 4 NJEx2011 AOV: analysis of variance of acoustic features between different levels of stress: higher (H), medium (M) and neutral speech (N). Signif. codes: ***$<$ 0.001, **$<$ 0.01, *$<$ 0.05, • $<$ 0.1, $-$ $<$ 1. Preliminary classification results are presented for the different sets.

greater for medium than for neutral and the spectral entropy values in average are smaller for medium than for neutral, which indicates an increase in the proportion of energy in higher frequencies. According to [10] these are characteristics of cognitive load or stress due to task load/engagement. On the other hand, significantly different features between higher stress and medium and neutral speech data (H/(M&N)) are mainly f0 and TEO features. In average f0 for higher stress is greater than f0 for medium and neutral data together. Taking into account the studies in [9, 10], we can conclude that indeed our annotated higher stress corresponds to physical or emotional stress.

Preliminary classification results of neutral speech and two levels of stressed speech are presented in Table 4. Three classifiers are trained with different sets of features, one for classifying three classes H/M/N and two for classifying two classes M/N and H/(M&N). Since the data is very unbalanced a weighted support vector machine (SVM) classifier is used; weight values are determined by the proportion of data in each class. 20 repetitions of stratified sampling are performed, where 2/3 of the data in each class is randomly selected to train the models and the other 1/3 is used for testing. The preliminary results indicate that the detection of higher and medium levels of stress is improved when the classifiers are trained with different

sets of features. For example the detection of higher stress respect to medium and neutral improved from 43% to 71% when using the H/(M&N) classifier.

5 Conclusions and future work

In this paper we have presented on-going work on analysis 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. We were able to annotate and identify the acoustic correlates of two types of stress on the recordings: physical stress and cognitive load. Traditional prosody features and sub-band acoustic features probed to be robust to discriminate among the different types of stress and neutral data. Our future work is to design appropriate classifiers of stress for the USAR domain that can cope with the very unbalanced data; when designing the classifiers we will take into account that the acoustic correlates of the two types of stress are very different, so the classifier/detector of physical stress should not be trained with the same features as the classifier/detector of cognitive load.

Acknowledgements The work reported in this paper has received funding from the EU-FP7 ICT 247870 NIFTi project. We would like to thank Holmer Hemsen for assistance with data annotation.

References

1. Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., Metze, F., Huber, R.: Detecting real life anger. In: IEEE Int. Conf. ICASSP. Taipei, Taiwan (2009)
2. Charfuelan, M., Schröder, M.: The vocal effort of dominance in scenario meetings. In: Interspeech. Florence, Italy (2011)
3. Hansen, J., Patil, S.: Speech under stress: Analysis, modeling and recognition. In: Speaker Classification I, *Lecture Notes in Computer Science*, vol. 4343, pp. 108–137. Springer (2007)
4. Hansen, J.H.L., Bou-Ghazale, S.E.: Getting started with susas: a speech under simulated and actual stress database. In: Eurospeech. Rhodes, Greece (1997)
5. Jameson, A., Kiefer, J., Müller, C., Gromann-Hutter, B., Wittig, F., Rummer, R.: Assessment of a user's time pressure and cognitive load on the basis of features of speech. In: Resource-Adaptive Cognitive Processes, Cognitive Technologies. Springer (2010)
6. Kruijff, G.: Proceedings of NJEx 2011, NID 2011 (2012). Unpublished report
7. Looije, R., te Brake, G., Neerincx, M.: Geo-collaboration under stress. In: Workshop on Mobile HCI for Emergencies. Singapore (2007)
8. Misra, H., Ikbal, S., Sivadas, S., Bourlard, H.: Multi-resolution spectral entropy feature for robust ASR. In: IEEE Int. Conf. ICASSP. Philadelphia, PA, USA (2005)
9. Patil, S.A., Hansen, J.H.L.: Detection of speech under physical stress: Model development, sensor selection, and feature fusion". In: Interspeech. Brisbane, Australia (2008)
10. Scherer, K.R., Grandjean, D., Johnstone, T., Klasmeyer, G., Bänziger, T.: Acoustic correlates of task load and stress. In: ICSLP2002 - Interspeech 2002. Denver, USA (2002)
11. Sjölander, K.: The Snack Sound Toolkit. http://www.speech.kth.se/snack (2012)
12. Zhou, G., Hansen, J.H.L., Kaiser, J.F.: Nonlinear feature based classification of speech under stress. IEEE Transactions on Speech and Audio Processing **9**(3), 201–216 (2001)