Analysis of speech under stress and cognitive load in USAR operations

Marcela Charfuelan, Geert-Jan Kruijff

DFKI GmbH, Language Technology Laboratory, Saarbrücken and Berlin, Germany

{marcela.charfuelan, gj}@dfki.de

Natural Human-Robot Cooperation in Dynamic Environments

Introduction

- **On-going work on analysis of speech under stress and cognitive load in speech recordings of Urban Search and Rescue (USAR) training operations:**
  1. We analyse human communication between team members on the field and members in the control command.
  2. We were able to annotate and identify the acoustic correlates of two types of stress on the recordings: **physical stress** and **cognitive load**.
  3. Traditional prosody features and acoustic features extracted at sub-band level probed to be robust to discriminate speech in very noisy situations.

Data and Method

Data

- Recordings of the NIFTi Join Exercises 2011 on human-robot-teaming (NJEx2011)
- 11 sessions (missions) where different team players (persons) participate in each session

Method

- Sessions were segmented by utterances:

	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.

NJEx2011 distribution of turns per day and speaker.

- Utterances were annotated according to three levels:
  - **Neutral** (level 1) : unstress, normal or neutral speech, happy, relax;
  - **Medium** (level 2): stress, speech is nervous, there is tension in the voice, more speed, there are hesitations;
  - **Higher** (level 3): high stress, there are shouts, anger, despair.

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%

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

- For analysis of stress we consider the utterance where the two annotators agree:

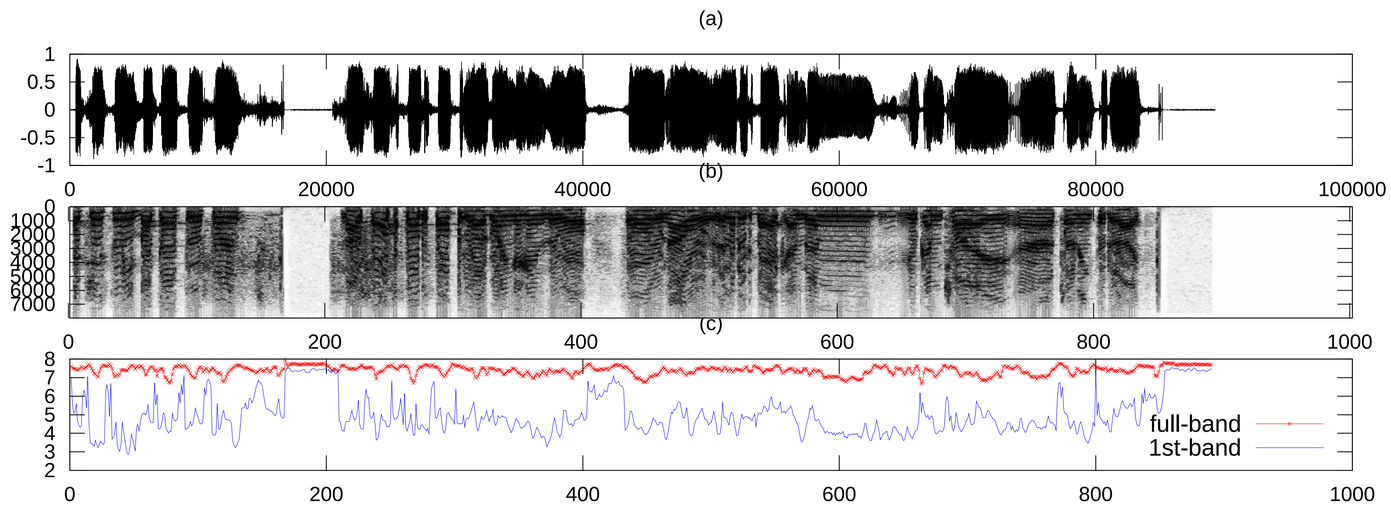
Stress level	Neutral	Medium	Higher	Total turns
Neutral	**1658**	287	2	1947
Medium	118	**226**	14	358
Higher	3	23	**24**	50
Total turns	1779	536	40	2355

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

- Acoustic measures are extracted from each utterance at frame and utterance level.
- ANOVA is performed among different sets, to identify acoustic correlates of each type of annotated stress.
- Preliminar Classification results using Support Vector Machine (SVM) are performed to discriminate different sets.

Data collection and annotation

NJEx2011 USAR training sessions: The FDDO ELW3 mobile command post, the Red Building, and the staff room in the ELW3.



Speech recordings of the NJEx2011 USAR training sessions: Speech wave (a), Spectrum of an utterance (b), and Spectral entropy calculated for the full-band signal (red) and the first band (0-1kHz) filtered signal (blue) (c).

Acoustic correlates of higher and medium stress types

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

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.**

Acoustic measures

- **Full band features:**
  - (a) Standard prosodic features: fundamental frequency (f0), duration, voicing rate, log power etc.
- **Sub-band features:**
  - (b) Teager Energy Operator - Autocorrelation Envelope (TEO-AutoEnv): TEO operator $\Psi[s(n)] = s^2(n) - s(n+1)s(n-1)$
  - (c) Voicing strengths (STR): correlation coefficient of s and delay t is defined by $c_t = \frac{\sum_{n=0}^{N-1} s(n)s(n+1)}{\sqrt{\sum_{n=0}^{N-1} s^2(n) \sum_{n=0}^{N-1} s^2(n+t)}}$
  - (d) Spectral entropy (SPE): $H(x) = -\sum_{x \in X} x_i * log_2 x_i$ where $x_i = \frac{X_i}{\sum_{i=1}^{N} X_i}$ $i = 1 : N$ and X_i is the spectrum of s

Conclusions

- 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.
- Our future work is to design appropriate classifiers of stress for the USAR domain that can cope with the very unbalanced data.