INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The SRI CLEO Speaker-State Corpus

Andreas Kathol, Elizabeth Shriberg, Massimilano de Zambotti

SRI International, Menlo Park, CA, USA
{andreas.kathol,elizabeth.shriberg,massimiliano.dezambotti@sri.com}@sri.com

Abstract

We introduce the SRI CLEO (Conversational Language about Everyday Objects) Speaker-State Corpus of speech, video, and biosignals. The goal of the corpus is providing insight on the speech and physiological changes resulting from subtle, context-based influences on affect and cognition. Speakers were prompted by collections of pictures of neutral everyday objects and were instructed to provide speech related to any subset of the objects for a preset period of time (120 or 180 seconds depending on task).

The corpus provides signals for 43 speakers under four different speaker-state conditions: (1) neutral and emotionally charged audiovisual background; (2) cognitive load; (3) time pressure; and (4) various acted emotions. Unlike previous studies that have linked speaker state to the content of the speaking task itself, the CLEO prompts remain largely pragmatically, semantically, and affectively neutral across all conditions. This framework enables for more direct comparisons across both conditions and speakers. The corpus also includes more traditional speaker tasks involving reading and free-form reporting of neutral and emotionally charged content. The explored biosignals include skin conductance, respiration, blood pressure, and ECG. The corpus is in the final stages of processing and will be made available to the research community.

Index Terms: Speech corpora, psychophysiology, autonomic nervous system, speech features, emotion

1. Introduction

The literature on speaker state is sizeable and growing (e.g., [1], [2], [3], [4], [5], [6], [7], [8], [9]). Yet there are few resources for the controlled study of how features in continuous speech vary with subtle influences on physiological state. Acted data or short responses do not tap the same processes as continuous spontaneous speech production. Speech in the wild, while ecologically valid, brings inherent variability on a wide range of dimensions (acoustic, semantic, pragmatic, affective, speaker-based) that make it difficult to assess specific questions about the relationship between changes in a speaker's speech features and speaker physiology.

To address this need we designed a speaking task that is simple enough for subjects to perform and sustain, but that provides enough richness and naturalness to use repeatedly across conditions. Speakers were prompted by collections of pictures of neutral everyday objects and were instructed to provide spoken descriptions related to any subset of the objects for a preset period of time (120 or 180 seconds). We refer to this as the "CLEO" (conversational language about everyday objects) task.

The CLEO task and corpus design allow for the controlled study of spontaneous speech and physiological measures while offering some distinct advantages. Unlike reading, participants must think about what to say. But, unlike spontaneous autobiographical recall elicitions, the discourse focuses on objects controlled by the experimenter that are extensible and semantically neutral. The task also produced lengthy speech samples. Most subjects were able to sustain their talk for more than a minute. This yields longer samples for study. It also assures that speech samples cannot be pre-planned.

Furthermore, we utilized a number of ways to induce different physiological states in the participants while they were performing the speech tasks. These include manipulations of their emotional state by exposing them to calm or emotionally stressful social situations, imposing an additional cognitive task diverting their attention, and imposing time constraints on their performance. These were chosen as representative of the kind of stress-inducting triggers participants are likely to encounter during their day-to-day interactions.

The corpus includes a number of biosignals as well as aligned high-quality audio and video recordings. Additionally, participants were asked to fill out an anonymized survey with basic health information at the time of the recording.

As noted in Section 6, the corpus is being made available to the community for research purposes.

2. Speech and biosignals

The sessions were conducted at SRI's Sleep Laboratory, which is equipped with recording devices for biosignals. Participants were seated at a small table and instrumented with sensors to collect ECG, respiration, blood pressure, and skin conductance signals. The biosignals were measured with the following devices:

- Thoracic Piezo Grael Rip Bands recorded the breathing signal through the ProFusion neXus platform using Grael amplifiers (Compumedics, Abbotsford, Victoria, Australia) with a sample rate of 64 Hz.
- A Portapres Model-2 measured systolic (SBP, mmHg), diastolic (DBP, mmHg), and mean blood pressure (MBP, mmHg) from the participants' middle finger of the non-dominant hand using a cuff that inflates and deflates continuously.
- A BioDerm Skin Conductance Meter measured skin conductance level (SCL, μm hos; the reciprocal of skin resistance) by using a constant 0.5 volt circuit between Meditrace Ag/AgCl surface spot electrodes attached to the thenar and hypothenar eminences of the non-dominant palm.

http://dx.doi.org/10.21437/Interspeech.2016-1141

- ECG recordings were performed using Meditrace Ag/AgCl surface spot electrodes placed in a modified Lead II Einthoven configuration through the ProFusion neXus platform using Grael amplifiers with a sample rate of 512 Hz.
- Additionally, Fitbit Charge HR wristbands were used to record heartbeat.

The participant interactions with the experimental laptop were recorded via Camtasia using the built-in Apple MacBook camera for the video channel and a Jabra GN2000 close-talking microphone for the audio channel. These audio-visual recordings were aligned with a separate low-bandwidth (2 KHz) audio recording that was also controlled by the ProFusion software in alignment with the biosignal inputs.

3. Participant statistics

A total of 42 participants were recruited, most of them employees of SRI International or their relatives. The session length was 60 minutes for participants 1–16 and 90 minutes for participants 17–43. Compensation was $50 for 60-minute sessions and $75 for 90-minute sessions. The participant gender distribution was 27 females and 16 males. The age range was 18 to 84 year, with an average age of 45.02 years.

Following the session, the participants were given a questionnaire to establish basic variables, including ethnicity; age; weight; native language; recent medical history; the 36-Item Short Form Survey Instrument (RAND 36-Item Health Survey 1.0); and the Profile of Mood States (POMS) survey. The latter questions participants for their emotional state in the preceding week along a four-part scale for descriptors such as "friendly," "tense," "angry," or "worn out."

4. CLEO picture-description tasks

The task prompts were given to the participants by means of a partially self-timed PowerPoint presentation. After initial instructions, the experimenters did not interact with the participants until the end of the session. When ready, each participant started a speaking task by pushing a button, which launched a 30-second gray screen, followed by the speaking task itself (120 or 180 seconds), and another 30-second display of the gray screen. A summary of the CLEO tasks, together with speaking duration (in seconds) is provided in Table 1.

The majority of the speaking tasks involved picture descriptions. The participants were given a grid of pictures of everyday objects that they were likely familiar with (cf. umbrella, bicycle, aquarium, record player in Figure 1). They were told to "to talk about these items, specifically as if in conversation with someone who has no prior knowledge of these items." Each of the different picture grids also repeated the relevant prompts, including: "What is the item? What does it look like? How is it used? What is its purpose? Do you have any stories from your own personal history involving that item?"

The participants were told to fill the time between gray screens with talk about these objects in any order, skipping any object or coming back to any object previously discussed. Participants were given a prerecorded example of the description task as well as a one-minute practice without any affective condition.

Table 1: *Summary of CLEO tasks and speaking duration.*

Task Type	Task Name	Duration
Picture description A/V background	Practice	60
	Tropical beach A/V	120
	Conversation A/V	120
	Marital argument A	180
	Teenager bullying A/V	180
	Custody dispute A/V	180
Picture description Cognitive load	Cognitive load	180
Picture description Time pressure	Time pressure	2x60
Free-form autobiographical recall	Trip to store	120
	Angry event	120
Reading	Child abuse	self-timed
	Bike-to-work	self-timed
Picture description Acted emotion	Neutral	120
	Furious	120
	Serene/relaxed	120
	Happy/excited	120
	Depressed	120
	Neutral	120

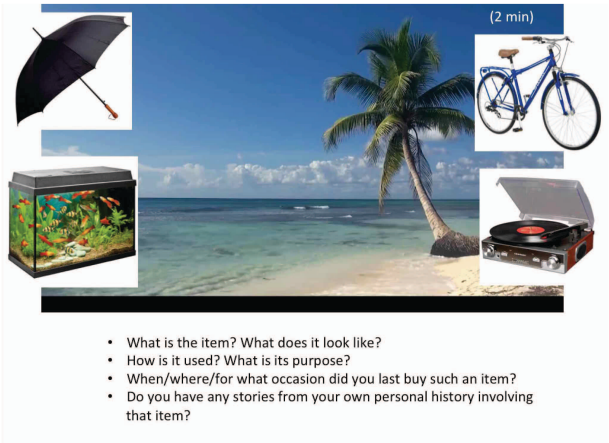

Figure 1: *Picture-description task with calm audio-visual background*

4.1. Audio-visual background

Five speaking tasks involved affective conditions with audio-visual background. That is, the picture grid was shown while an audio clip or movie was played in the background. These consisted of two neutral/calming audio-visual backgrounds: (1) a tropical beach scene (see Figure 1) with wave sounds and (2) calm conversation about a noncontroversial topic.

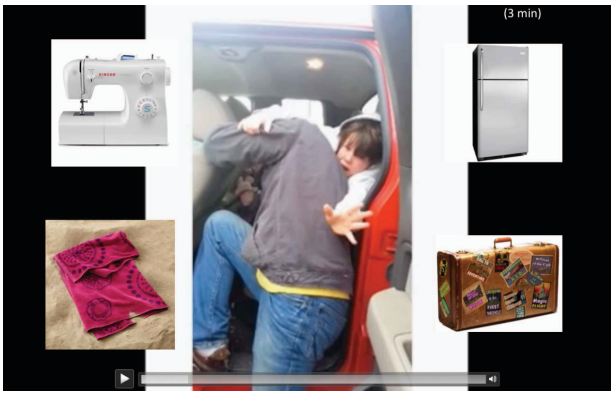

Figure 2: *Screenshot of picture-description task with evocative audio-visual background (child in custody dispute).*

The speaking time for each was 120 seconds. The second set of affective conditions was chosen to elicit a strong emotional response from the participants and consisted of (1) a recording of a marital dispute (audio only), (2) footage of teenager being bullied and physically assaulted by peer, and (3) footage of an upset child being removed from his mother in a custody dispute (see Figure 2). The speaking time for this second set was 180 seconds each.

4.2. Cognitive load

In the cognitive load condition, the picture description was paired with the task of keeping track of visual stimuli that

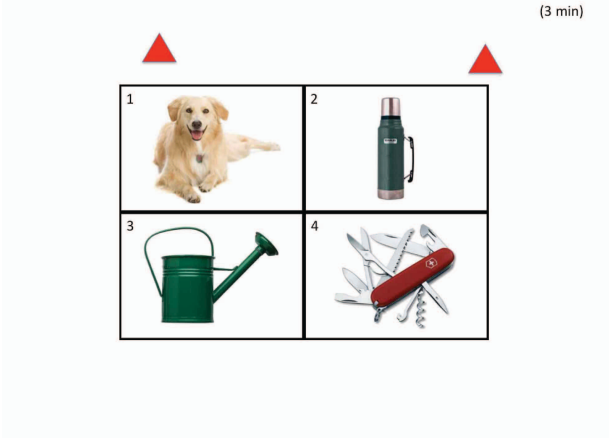

Figure 3: *Picture-description task under cognitive load.*

occur briefly (< .5 seconds) in the background. Specifically, the participants were asked to count the number of pairs of red triangles among all pairs of objects, which also included non-matched triangle/rectangle combinations (see Figure 3). The speaking time for the cognitive load task was 180 seconds.

4.3. Time pressure

In the time pressure condition, the participants were shown a panel of nine pictures and told to describe as many of them as possible in one minute (see Figure 4). The remaining speaking time was indicated by means of a clock icon. After the first

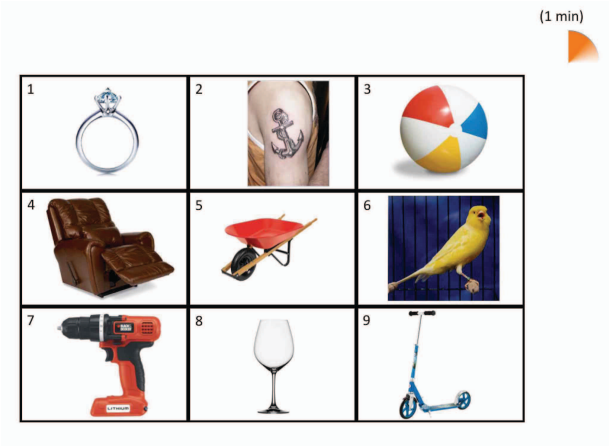

Figure 4: *Screenshot of picture-description task under time pressure.*

minute, another panel of nine pictures was shown, again with the task of describing as many as possible in one minute.

4.4. Acted emotions

For a subset of 27 out of the total 43 participants, a number of additional picture-description tasks were added that asked them to perform the speaking task in a tone of voice that reflects a particular emotional state (see Figure 5). The four states of interest were *furious*, *serene/relaxed*, *happy/excited*, and *depressed*.[1] In each state, a different panel of four pictures was presented with a speaking time of 120 seconds. Preceding and following the four explicit emotional states were two more speaking tasks with a "neutral" tone of voice.

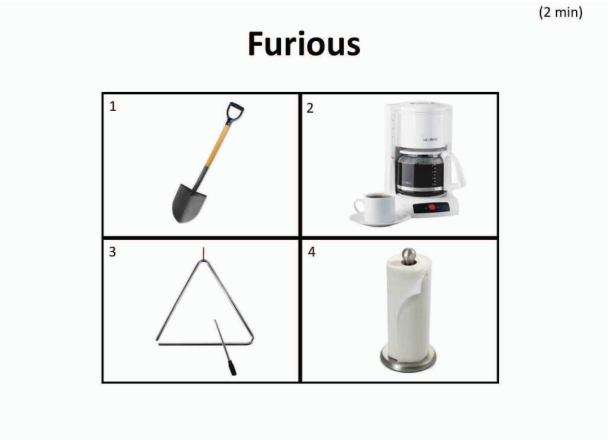

Figure 5: *Screenshot of picture-description task with acted emotion (furious).*

[1] This task somewhat deviates from other ones with respect to naturalness in that some participants found it difficult to willfully modulate their style of speaking if the objects to describe did not have any intrinsic emotional value for them.

5. Additional comparative tasks

As noted earlier, the CLEO picture-description task was designed specifically so that the speech content would be independent of the affective condition. To facilitate comparison to prior work, we additionally included two types of tasks that are typically used in the literature on affect, mental health, and trauma narratives. Read speech (cf. [10]) is often the first type of audio collected for health-based research, but retelling of an episode from personal experience (cf. [11]) is a commonly used task to evoke stronger emotions ([12]). However, the latter introduces complications in the form of variability across speakers and contexts. In our design, we included both a reading and autobiographical recall task and paired each one with a neutral control task that was designed to be emotionally neutral.

5.1. Reading

Participants were given two short newspaper excerpts to read, one with emotionally evocative content (child abuse), and one without (bike-to-work day).

5.2. Free-form autobiographic recall

Participants were asked to recount two episodes from their own personal experience, one with emotionally evocative content ("a time when you felt really angry or annoyed about something or someone"), and one without (a recent visit to the grocery store). The speaking time was 120 seconds.

6. Summary and availability

The SRI CLEO corpus uses a novel speaking task and design, that allows for the controlled examination of lengthy natural speech samples under different speaker state conditions. The elicitation method results in subtle effects on speaker state that support analysis of the relationship between speech features and biosignals either across or within states. The corpus also contains a number of comparative data conditions as well as acted speech conditions.

By using a variety of background conditions for the speech tasks, we ultimately hope to gain better understanding in how speech features are reflective of particular user states, as defined in physiologically measurable terms, especially if these have potential long-term negative effects on a person's well-being.

The corpus processing is in its final stages. Many signals are currently available for research purposes. These parts of the corpus can be obtained by directly emailing the authors. The corpus will be also be submitted for general distribution in 2016-2017.

7. Acknowledgements

The collection reported on here was conducted under NSF Early Grant for Exploratory Research (EAGER) #1449202 "A Corpus of Aligned Speech and ANS Sensor Data."

8. References

[1] Banse, R. and Scherer, K. (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, Vol.70, 614-636.

[2] Bänziger, T. and Scherer, K. (2005). The role of intonation in emotional expression. Speech Communication, , Vol.46, 252-267.

[3] Busso, C., Mariooryad, S., Metallinou, A., and Narayanan, S. (2013). Iterative feature normalization scheme for automatic emotion detection from speech. IEEE Transactions on Affective Computing, 4(4), 386-397.

[4] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G. (2001). Emotion recognition in human-computer interaction. Signal Processing Magazine, IEEE,18(1), 32-80.

[5] Hansen, J. and Patil, S. (2007). Speech under stress: Analysis, modeling and recognition. In: Speaker Classification I, Lecture Notes in Computer Science, vol. 4343, pp. 108–137. Springer.

[6] Lee, C.-C., Mower, E., Busso, C., Lee, S., and Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach, Speech Communication, vol. 53, no. 9-10, pp. 1162-1171.

[7] Mower Provost, E. (2013). Identifying Salient Sub-Utterance Emotion Dynamics Using Flexible Units and Estimates of Affective Flow. Proc. ICASSP.

[8] Picard, R., Vyzas, E. and Healey, J. (2001). Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(10), 1175-1191.

[9] Schuller, B. and Batliner, A. (2013). Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. Wiley.

[10] Graesser, A. C. and D'Mello, S. (2012). Moment-To-Moment Emotions During Reading. The Reading Teacher. vol. 66, no 3, pp. 238–242, November 2012.

[11] Mills C. and D'Mello, S. (2014) On the Validity of the Autobiographical Emotional Memory Task for Emotion Induction. PLoS ONE 9(4): e95837. doi:10.1371/ journal.pone.0095837.

[12] Mitra, V. and Shriberg, E. (2015) Effects of Feature Type, Learning Algorithm and Speaking Style for Depression Detection from Speech, Proc. ICASSP, pp. 4774–4778.