The D-ANS Corpus: the Dublin-Autonomous Nervous System Corpus of Biosignal and Multimodal Recordings of Conversational Speech

Shannon Hennig, Ryad Chellali, Nick Campbell
Istituto Italiano di Tecnologia / Trinity College
Genova, Italy / Dublin, Ireland
Email: hennig.iit@gmail.com, ryad.chellali@iit.it, nick@tcd.ie

Abstract

Biosignals, such as electrodermal activity (EDA) and heart rate, are increasingly being considered as potential data sources to provide information about the temporal fluctuations in affective experience during human interaction. This paper describes an English-speaking, multiple session corpus of small groups of people engaged in informal, unscripted conversation while wearing wireless, wrist-based EDA sensors. Additionally, one participant per recording session wore a heart rate monitor. This corpus was collected in order to observe potential interactions between various social and communicative phenomena and the temporal dynamics of the recorded biosignals. Here we describe the communicative context, technical set-up, synchronization process, and challenges in collecting and utilizing such data. We describe the segmentation and annotations to date, including laughter annotations, and how the research community can access and collaborate on this corpus now and in the future. We believe this corpus is particularly relevant to researchers interested in unscripted social conversation as well as to researchers with a specific interest in observing the dynamics of biosignals during informal social conversation rich with examples of laughter, conversational turn-taking, and non-task-based interaction.

Keywords: biosignals, unscripted conversation, social interaction

1. Introduction

There is a growing number of multimodal corpora available to researchers interested in spoken and nonverbal communication during social interaction. Such data is important for fully understanding the complexities of how humans engage with each other, share information, mask information, and connect socially.

Existing corpora vary in terms of participant interactivity and number of signals that are captured. The TIMIT corpus (Garofolo et al., 1993) of read speech is an example without interaction. The CALLHOME corpus (Canavan et al., 1997) of 120 recorded telephone conversations is more interactive, but is audio only. More interactive is the AMI meeting corpus (McCown et al., 2005) which includes audio, video, and slides from task-based interactions during staged meetings (conducted between actors) as well as real design meetings. More interactive still is the D64 corpus (Oertel et al., 2013), in which several people agreed to be recorded in an apartment setting. This allowed the capture of free flowing interaction including the formation (and break up) of various sized conversational groupings, free movement about the space, drinking of tea and wine, and the type of unscripted social chat that is ubiquitous in human society.

1.1. Rationale for Capturing biosignals

Biosignals, while challenging to interpret, are thought to provide insights into difficult to observe internal states (Dawson et al., 2007), such as changes in cognitive load, affect, and arousal. Changes in speech behavior are thought to be associated with changes in emotional state, particularly physiological arousal (Juslin & Laukka, 2003)

1.1.1. Experimental and classification studies

Experimental work from studies on anxiety during public speaking (Elfering and Grebner, 2011; Gerlach et al., 2001), communication during stress (Hansen and Bou-Ghazale, 1997), memory recall of spoken narratives (MacWhinney et al., 1982), and clinical interviews during investigations on therapeutic rapport (Marci et al., 2007) suggest possible links between biosignals and social communication behaviors.

Specifically with regard to empathy, MacWhinney, et al., (1982) report that EDA recorded during a conversation was related to degree of involvement between participants in the moment. Gallo et al's (2000) findings suggest that social context plays a role on EDA responses during communication. Guastello et al., (2006) found that during 20 minute discussions between college students, empathy appeared to interact in a nonlinear fashion with EDA over time. Further, Marci et al.'s reviewed the relevant literature and concluded that EDA appears to be associated with emotional and empathic responsiveness more consistently than other physiological measures (2007).

Additionally, work on the classification of emotional and affective state from speech, communication behavior, and biosignals (Calvo and D'Mello, 2010; Kim, 2007) suggests potential links between biosignals and affective states during speech.

In general, existing experimental work using biosignals typically has reported differences in average electrodermal and/or cardiac activity between conditions without examining dynamics of how these signals change over time. To date very little is known about how such fluctuations in biosignals vary during social, interactive, spontaneous, free-flowing conversation. Even less is known about the dynamics of these signals over extended periods of time in natural contexts.

1.2. Motivation for D-ANS Corpus

As scientific recognition of the importance of nonverbal behavior increased, coupled with advancements in recording technology, a transition occurred from primarily audio-only corpora to multimodal corpora containing audio, video, and other signals (i.e., motion capture). Similarly, questions about the affective and internal states of participants during social interaction has increased interest in utilizing wireless, noninvasive, wearable sensors to safely record biosignal.

The D-ANS corpus was recorded to provide insights into how changes in biosignals, specifically heart rate and electrodermal activity, may relate to multimodal communication during spontaneous conversation. While databases of physiological reactions to various stimuli have been shared (e.g., DEAP, Koelstra et al., 2012) and biosignals have been used for affect classification (see Calvo & D'Mello, 2010 for a review), there is limited shared data that allows for the observation of biosignals during conversation in order to ground and inform future work in this area. To the best of our knowledge D-ANS is the first shared corpus of multimodal corpus of conversation in which noninvasive biosignals were worn and it was collected to allow researchers the opportunity to directly observe these signals in context.

2. Corpus Collection

2.1. Setting

A comfortable, informal sitting area was used that consisted of a 3 person sofa and a stuffed arm chair arranged around a coffee table (see figure 1). The course of conversation was allowed to unfold naturally without external manipulation.

2.2. Task

We are particularly interested in the dynamics of social phenomena, such as engagement, interpersonal support and rapport and synchrony, that are known to be influenced by the social context.

For the purpose of informed consent, basic information was provided regarding the biosensors and recording equipment, but no direction was given regarding the structure or objective of the interaction other than to sit down and "chat."

Figure 1: The informal conversational space used with seat positions and microphones labeled.

2.3. Participants

Five adults participated over the course of three days. The corpus includes interactions between 2, 3 and 4 people, all of whom had previously conversed with each other in a social and professional capacity.

Specifically, on day 1, two people interacted. On day 2, three people interacted and were later joined by a 4th for the final hour of recording. On Day 3, two people began the interaction and were joined by a third for the final portion of the recording.

As outlined in table 1, there were 3 men and two women. Four were native English speakers and one French speaker with near native English fluency. A variety of English accents were represented including American, Irish, and British.

Subject	*Gender*	*Accent*	*Present*	*Heart rate*
1	Male	British English	Day 1, seat 1 Day 2, seat 1	Day 2
2	Female	American English	Day 1, seat 2 Day 2, seat 3 Day 3, seat 2	No
3	Male	Irish English	Day 2, seat 2	No
4	Male	Irish English	Day 2, seat 4 Day 3, seat 3	Day 3
5	Female	French	Day 3, seat 1	No

Table 1: Breakdown of participants by recording day.

2.4. Description of Recording Days

In naturally occurring social interactions, groups naturally self-organize with individuals entering and leaving conversational groups. Given this, participants were allowed to freely come and go from the sessions allowing such transitions to be included in the corpus.

2.4.1. Day One

Day one was the warm-up day and was effectively a test of the equipment and recording set-up. This was the

most task-based of the days given that the participants were discussing the motivations for and planning how to execute the corpus collection. This session lasted approximately one hour, was conducted between two researchers on the project. It had a similar tone to a typical work meeting between colleagues. This data is being held in reserve for future comparison with the more socially interactive day 2 and day 3 data.

2.4.2. Day Two

Day two was the longest recording session with over 3 hours of continuous recording. The first two hours were between three participants and then a fourth participant joined the conversation for the final hour. This day had a more social, informal feel than the first day with people telling stories about personal interests and past experiences while trying to find shared interests. As commonly occurs during free conversation, many topics were unsurprising given the specific shared background of the participants (i.e., in this case, academics discussing grading schemes, universities and cities they had previous worked at, and conference travel), however most topics were arguably less predictable (e.g., stories of being mugged, destruction of specific Irish landmarks by vandals, taking photographs of manhole covers during one's travels, etc.). Such a mix of predictable and unpredictable topics is not uncommon during everyday, social conversation.

2.4.3. Day Three

Day three consists of conversation between speaker 5, who spoke English as a second language, and participants 2 and 4. In contrast to the first two days, on day three all participants were peers, both in terms of age and professional position. This recording is approximately one hour long and topics included religion, roller-coasters, and experiences growing up in different countries. The third participant joined the group after the conversation began and had to briefly excuse herself to take a phone call. This resulted in two sections of dyad interaction and two sections of triad conversation.

2.5. Audio and Video Equipment

Audio was captured using five high quality microphones. This included 2 Sennheiser shotgun microphones (MKH 60p) (one positioned near the arm chair and one on the right side of the sofa), a third microphone (Sennheisser MKH30 P48 cardiod) recorded the their participant's voice from behind the sofa, and fourth boom microphone was placed above and in front of the fourth seat. These four audio streams were recorded at 48kHz on a Marantz MOTU four channel digital sampler. Additionally, a Roland R09 mk II portable field recorder was positioned approximately equidistant from all participants on the center coffee table. The location of the microphones relative to seats is illustrated in figure 1.

Three video camera angles were recorded, as illustrated in figure 2.

Figure 2: Three camera angles: global overview on top, chair cam bottom left, and sofa cam bottom right

A global overview was provided by a high quality Logitech C930 HD webcam positioned approximately a meter above the participants' heads. Two Sony HDR XR500 digital video cameras were positioned to provide a frontal camera angle of the sofa and chair respectively. These two camera angles overlapped such that the middle person was captured on both cameras. This allowed interactions between the two people on the sofa or the two people seated closest to the angle between the chair and sofa to each be captured in a single camera angle. In the sections of the corpus in which 4 people were interacting, the fourth person was seated in a chair next to the sofa, and unfortunately frequently leaned back and out of frame.

2.6. Biosignal Sensors

Biosignals are thought to be an indirect measure of changes in the autonomic nervous system (ANS) and recent technological advances allow both EDA and heart rate to be safely be recorded from cordless, noninvasive, wearable sensors. In the past, such recordings required participants to be 'wired up' and be physically tethered to a computer, interfering with their ability to freely gesture during communication.

2.6.1. Electrodermal Activity

Here we focus primarily on Electrodermal Activity (EDA, also known as galvanic skin response and skin conductance) which is a measure of how readily a small current of electricity passes across the skin. It is associated with activation of the sympathetic branch of the ANS and is correlated with increases in physiological arousal (Dawson et al., 2007). Changes in EDA also associated with changes in attention, perception, problem-solving, movement, and emotion (Calvo & D'Mello, 2010; Dawson et al., 2007).

To record EDA, six Q sensors (www.affectiva.com, since discontinued, 32 Hz) were worn on the under-side of each participants' wrist. The exception was on Day 2 when four participants were recorded. For this day, only two people wore two sensors, the other two participants wore one.

The Q sensors also recorded 3 degrees of acceleration, capturing capturing wrist movements. These sensors were selected for their non-invasive form-factor; however this comes with a tradeoff of capturing a less fine-grain data than can be obtained with traditional sensors that record from the palm or finger tips.

2.6.2. Heart rate

Heart rate was also recorded using a mass-market Polar CS600 heart rate monitor which recorded R-R heart rate in milliseconds from a Polar chest strap (www.polar.fi). Heart rate is associated with changes in both the sympathetic and parasympathetic branches of the nervous system (Bertson et al., 2007). Heart rate is known to vary with breathing (i.e. Rhythmic Sinus Arithymia, RSA), body position (i.e., supine or prone), stress, cognitive load, and other affective experiences (Bertson et al., 2007).

2.7. Synchronization

A custom synchronization procedure was used given that both the wireless Q-sensors and Polar heart rate monitor do not record data to a centralized system. To synchronize the EDA signals, the accelerometer data from the Q sensors were manually synchronized to movements on the videos. The heart rate was synchronized to the videos using by the event marker activations captured on the video that also recorded a time-stamp on the heart rate datafile. Video and audio were synchronized using the audio track of all files. All data files were cropped so they have a universal start time.

2.8. Annotations

At this stage, day three has been the most heavily annotated. Two annotators have marked silent and audible laughter (Gilmartin et al., 2013) using ELAN (Wittenburg et al., 2006) and Praat (Boersma, 2001). Additionally turn-taking dynamics (e.g., gaps, pauses, backchannels). Examples the temporal scale of the biosignals and a subset of laughter annotations is provided in figure 3.

It is hoped that by sharing this language resource that a larger collection of annotations of potentially relevant phenomena can be amassed in order to collectively advance our understanding in this area.

2.9. Challenges Encountered

As anticipated, a small number of sensor malfunctions occurred. On day three only three microphones recorded correctly: two boom microphones and the portable recorder positioned on coffee table. On all days, only one heart rate monitor correctly recorded. Occasionally, one Q sensor per participant was significantly noisy suggesting a poor sensor connection (e.g., participant 4's left hand on day 3); the impact of these artifacts is mitigated by the bilateral recordings, allowing relatively complete coverage of EDA for all participants from at least one wrist.

There are several caveats worth mentioning. A relatively small number of people was recorded and biosignals. EDA, in particular, is known to have significant interpersonal variability. Additionally, all participants were researchers, and while only one was specializing in

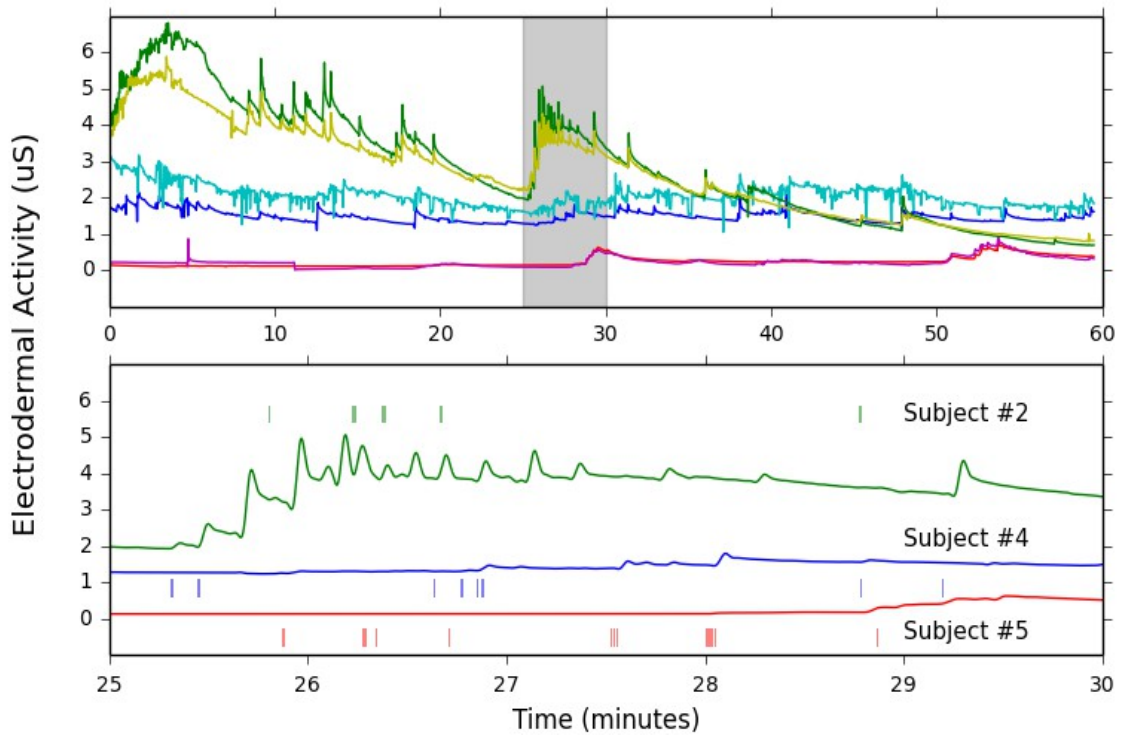

Figure 3: Electrodermal activity from participant 2, 4 and 5 on day 3. Top plot: entire record. Below: zoomed section of EDA from right hand of each person with each person's annotated audible laughter annotated in vertical bars.

biosignal data, their professional familiarity with multi-modal corpus work may have altered their behavior.

A final challenge is the lack of clearly established methods for analysing the temporal fluctuations of EDA, and to a less extent heart rate, in such a natural context. Both localized skin conductance responses (SCRs) and global changes in skin conductance level (SCL) are observable.

3. Distribution of Corpus

This corpus allows observation of the coarse temporal dynamics between biosignals and communication behaviors in the context of informal conversation. By gathering unscripted, spontaneous, interactive conversation in an informal and social context, it is possible to observe the moment to moment dynamics of biosignals over the course of a naturally flowing conversation.

Participants have agreed to share this data with the research community, provided that the details of the personal stories and identifying information (i.e., names, birth dates, etc.) caught on camera not be shared in any resulting publications or presentations and in general be treated as confidential. The annotations, media, and biosignal data will be shared on a website[1] along with sample video clips to allow any interested parties to have a sense of the type of interaction captured in this corpus.

The full corpus (3-5 audio files, 3 video files, biosignal csv file for each day of recording) will be made available for noncommercial research purposes to any interested researchers upon the return of signed release forms found on the website.

4. Conclusion

We believe this corpus has value to the larger community of researchers studying questions related social chat given the limited amount of data collected during informal, social conversation.

The specific value in the corpus is that it allows for the observation of biosignals, including their temporal properties, collected in an informal, conversational context. The resolution of the biosignals captured may not be appropriate for some machine learning approaches, however we have found reviewing this data invaluable particularly when considering published findings from controlled, lab-based experiments and classification studies and their hypothetical links to real-world phenomena. This corpus has the potential to inform future work as well as improve our understanding of how these signals vary (or not) with phenomena including rapport, turn-taking, entrainment, laughter, eye contact, turn-taking, cognitive load changes, and engagement.

1 Distribution website for data as well as a repository for shared annotations: http://www.speech-data.jp/d-ans/

5. Acknowledgments

This work was carried out in the Speech Communication Lab at Trinity College Dublin and was supported by the SFI FastNet (project 09/IN.1/1263). It was conducted as part of the first author's doctoral work, which was funded by Università degli Studi di Genova and the Instituto Italiano di Tecnologia. The work was co-funded as part of the Japanese Government KAKEN basic research into MOSAIC: Models of Spontaneous and Interactive Communication. We also thank the annotation efforts of Emer Gilmartin and Céline De Looze.

6. References

Bertson, G.G., Quigley, K.S. & Lozano, D. (2007). Cardiovascular Psychophysiology. In J.T. Cacioppo, L.G. Tassinary, and G.G. Berntson (Eds.), Handbook of Psychophysiology. New York: Cambridge University Press, pp. 182-210.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International, 5(9/10), pp. 341-345.

Calvo, R.A. & D'Mello, S. (2010). Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing, 1 (1), pp. 18-37.

Canavan, A. Graff, D. & Zipperlen, G. (1997). CALLHOME American English Speech. Philadelphia: Linguistic Data Consortium.

Dawson, M.E., Schell, A.M. & Dilion, D.L. (2007). The Electrodermal System. In J.T. Cacioppo, L.G. Tassinary, and G.G. Berntson (Eds.), Handbook of Psychophysiology. New York: Cambridge University Press, pp. 159–181.

Elfering, A. & Grebner, S. (2011). Ambulatory assessment of skin conductivity during first thesis presentation: lower self-confidence predicts prolonged stress response. Applied Psychophysiol Biofeedback, 36(2), pp. 93-99.

Gallo, L.C., Smith, T.W. & Kircher, J.C. (2000). Cardiovascular and electrodermal responses to support and provocation: Interpersonal methods in the study of psychophysiological reactivity. Psychophysiology, 37(3), pp. 289-301.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G. & Pallett, D.S. (1993). TIMIT acoustic-phonetic continuous speech corpus. Philadelphia: Linguistic Data consortium.

Gerlach, A.L., Wilhelm, F.H., Gruber, K., & Roth, W.T. (2001). Blushing and physiological arousability in social phobia. Journal of Abnormal Psychology. 110(2), pp. 247-258.

Gilmartin, E., Bonin, F., Vogel, C. & Campbell, N. (2013). Laughter and Topic Transition in Multiparty Conversation. In Proceedings of the Fourteenth SigDial Meeting on Discourse and Dialogue. Metz, France, pp. 304-308.

Guastello, S.J., Pincus, D. & Gunderson, P.R. (2006). Electrodermal arousal between participants in a conversation: Nonlinear dynamics and linkage effects.

Nonlinear Dynamics, Psychology, and Life Sciences 10(3), pp. 365-399.

Hansen, J.H.L., Bou-Ghazale, S., Sahar, E., Sarikaya, R. & Pellom, B. (1997). Getting started with SUSAS: A speech under simulated and actual stress database. In Proceedings of the Fifth biennial Eurospeech Conference, Rhodes, Greece, pp. 1743–1746.

Juslin, P.N. & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin, 129(5), pp. 770-814.

Kim, J. (2007). Bimodal emotion recognition using speech and physiological changes. Robust Speech Recognition and Understanding, pp. 265-280.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. & Patras, I. (2012). Deap: a database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing, 3(1), pp. 18-31.

MacWhinney, B., Keenan, J.M. & Reinke, P. (1982). The role of arousal in memory for conversation. Memory & Cognition, 10(4), pp 308-317.

Marci, C.D., Ham, J., Moran, E. & Orr, S.P. (2007). Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. The Journal of Nervous and Mental Disease, 195(2), pp. 103-111.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al. (2005). The AMI meeting corpus. In Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, 85.

Oertel, C., Cummins, F., Edlund, J., Wagner, P. & Campbell, N. (2013). D64: a corpus of richly recorded conversational interaction. Journal on Multimodal User Interfaces, 7(1-2), pp.19-28.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. & Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In Proceedings of the Fifth international conference on Language Resources and Evaluation, Genoa, Italy.