Annotating the TCD D-ANS Corpus – A Multimodal Multimedia Monolingual Biometric Corpus of Spoken Social Interaction

Nick Campbell[1(✉)] and Shannon Hennig[2]

[1] Speech Communication Lab, Trinity College Dublin, Dublin, Ireland
nick@tcd.ie
[2] Inclusive Communication Ltd., Wellington, New Zealand
shannon@inclusive-communication.co.nz

Abstract. This paper describes a recently created multimodal biometric corpus of spontaneous casual spoken interaction recorded at Trinity College Dublin, the University of Dublin, in Ireland, and currently being made available for wider dissemination. The paper focusses on the use of this corpus for training or learning about the needs and limitations of an interactive spoken dialogue interface for human-machine communication. Since the corpus is still very new and only recently released, the paper does not present research findings based on an analysis of the content but instead suggests methods and goals for annotating the material so that future researchers can use it to design more sensitive interfaces for speech synthesis in spoken dialogue systems. The paper is an extended version of an invited talk at the MA3HMI workshop.

Keywords: Spoken dialogue · Multimodal interaction · Biometric data · Capture & analysis · Interactive speech synthesis · Perceptual computing

1 Introduction

Human-computer interfaces for the general public are not new but they are rapidly becoming a key technology, as computing devices become smaller and more ubiquitous. Wearable or pocketable computers are now common, and the range of sensors they incorporate is growing at a rate we could not have predicted ten years ago.

Speech-based interaction with machines or knowledge-systems is no longer a dream but is now an everyday reality as the world of digital information is opening up to people-in-the-street, with young children now being exposed to smart devices with swipable and voice-activated interfaces perhaps even before they learn to use a pencil or pen.

'Perceptual computing' might now be a brand-name (of Intel) but it reflects the way that machines are becoming sensitive to humans in a more human-like way; incorporating gesture, tone-of-voice, speech and facial dynamics, and near-field interaction modalities as part of the basic operating system of a modern-day tablet or personal computer to enable new modes of interaction between people and machines (see Fig. 1).

R. Böck et al. (Eds.): MA3HMI 2014 Workshop, LNAI 8757, pp. 3–12, 2015.
DOI: 10.1007/978-3-319-15557-9_1

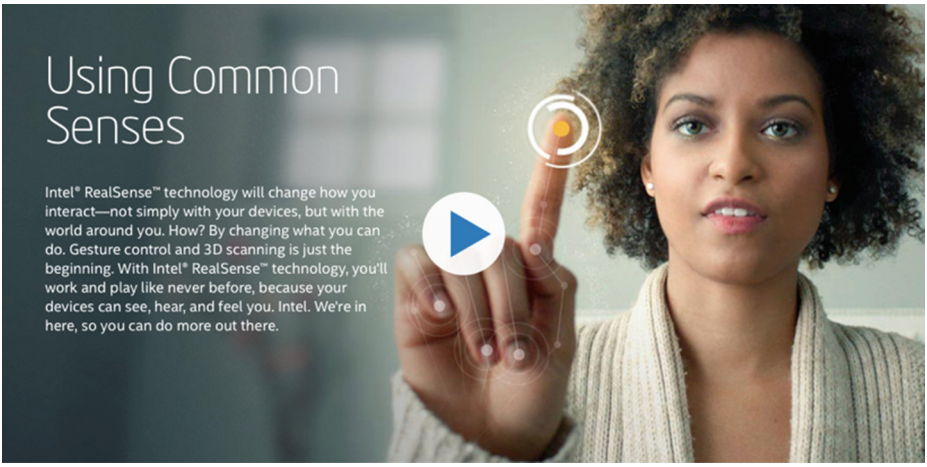

Fig. 1. From a recent Intel advert for Perceptual Computing. Note that the computer is aware of the shape of the hand (and probably facial expression) as well as being capable of speech and gesture processing

This technology which is with us now, will depend heavily on advances in natural-language processing and virtual-agent rendering to facilitate the natural forms of spoken interaction that are so characteristic of human social interaction. Devices will have to learn to read the signals that we commonly use to punctuate and inform our speech, and to 'read between the lines' of physical utterances and gestures to be able to infer the cognitive states and intentions that underly them.

It is therefore even more necessary that we should have a complete understanding of these processes so that advances can be made in the soft side of the technology to keep up with the rapid advances in computer hardware and memory capabilities. This inspired us to collect a corpus of normal everyday spoken interaction with not just audio and video recordings but also biometric sensors to provide a possibly more objective measure of the participant states and interactions in the course of a conversation. From this corpus we hope to learn how people signal their role(s) in a conversation so that computer interfaces might be better able to read those signals and act on them without the need for explicit commands.

The following sections briefly provide some background to the TCD D-ANS corpus [1], and mention some technological issues related to speech processing in human-computer interfaces before discussing the annotation requirements of the new corpus.

2 Serendipitous Liaisons

When Shannon Hennig came to visit our lab in November 2011, she brought with her a paper by Beukelman [2] of Nebraska whose analogy of 'right hand'

and 'left hand' messages (referring to the melody and harmony parts on a piano but with clear implications to human speech) seemed closely related to much of what we had discussed in earlier meetings about the different modes of speech activity in social contexts. It became the seed of an idea by which we determined to obtain measures of bilateral neural activity and learn something of its relation to task-based and chat-based, or formal and social modes of interaction in casual conversational speech.

Her interest at that time was in the candidacy of physiological measurements for implicit control of emotional speech synthesis. My earlier work on the development of expressive speech synthesis overlapped well with her ideas of morphological computation in natural speech (inspired by Pfeifer, Bongard and Grand, 2007 p. 96 [3]) whereby environmental triggers initiate Autonomous Nervous System (ANS) activity which results in physiological body changes such that vocal-tract constriction (for example) influences voice quality, inducing subtle changes in the quality of the speech signal that can be picked up by the listener who then infers para-linguistic or extra-linguistic information from that aspect of the signal in order to better parse the utterance in context.

We wanted to know if there are correlations between variations in physiological measures and vocal acoustic measures that we could use in (either or both) processing the input signal from humans engaging in conversation, and generating an equivalent signal rich in paralinguistic information for the synthesis of more natural-sounding machine-generated speech. Her then recent work with Autonomic Nervous System responses (measured using Affectiva's Q-Sensors [4]) in relation to stressful speaking situations convinced us that there was value in measuring and learning from similar responses in more casual informal social speech.

As detailed in [1], Q-sensors measure Electrodermal Activity (EDA, also known as galvanic skin response and skin conductance) which is how readily a small current of electricity passes across the skin. EDA is associated with activation of the sympathetic branch of the ANS and is correlated with increases in physiological arousal [5] Change in EDA is associated with changes in attention, perception, problem-solving, movement, and emotion [5,6].

3 Recording Setup

The Speech Communication Lab in the Centre for Language and Communication Studies at Trinity College Dublin has excellent facilities for multimodal recording of casual spoken interactions, both human-human and human-machine, so after the purchase of some extra Q-sensors, we were able to start recording with the help of friends and colleagues.

Consent forms were prepared and subjects informed of their rights to withdraw at any time, as well as being warned not to broach any particularly sensitive topics as their conversations were being recorded. All participants were familiar with the surroundings and equipment in the lab so none were in any way intimidated by being seated in the midst of microphones and cameras, though none except Shannon had any experience of wearing the wrist-watch-like Q-sensors.

Fig. 2. A scene from the D-ANS corpus (overhead webcam view) showing seating and microphone placement. Participants each wearing two Q-sensors

The first day of recordings was very much one of experiment. We needed to find ways to effectively and efficiently link a set of videos of people talking (and gesturing and moving about) with the accelometer data from their wrists during these videos. We needed to find optimal positions for camera placement and to find locations for microphones that would be able to pick-up fine vocal fluctuations while not being invasive or hampering the free movement of the speakers in any way.

We also discussed strategy and planned ways of maximising the variety of participants' speaking styles across various dimensions of formality, familiarity, and conviviality. The recordings from Day-1 are not part of the publicly-available corpus but do provide useful baseline measures from which we can compare the performance of the same subjects in the later recordings. They can however be made available to interested researchers upon request.

Figure 2 (from [1]) shows the layout of the studio corner where the recordings were made. Shannon is on the right and the first author on the left, with a colleague and friend from another Irish university in the middle. Participants were free to move around, and change seats. The relaxed atmosphere of the recordings can perhaps be seen from the poses of the participants. Freshly signed consent forms are visible on the table.

There are microphones in abundance, and cameras recording from several angles, but none of the equipment is intrusive in any physical way. Two clocks (radio syncronised) ensure that the output of all cameras can be roughly aligned. Shotgun microphones provide the main audio recordings but these are backed up by high-quality far-field microphones and a small portable stereo desktop

Fig. 3. A scene from the D-ANS corpus showing the different camera views.

recorder for simple fast navigation and backup. Lower-quality audio from the video cameras themselves can be used for accurate alignment of the videos.

Figure 3 shows a scene from Day-2 with the overview webcam display at the top and two high-definition video images from each of the working cameras below. The webcam, like the table-top Roland Edirol audio recorder, is primarily for backup and general overview processing for humans; the working cameras and the Sennheiser microphones are for more detailed machine-processing of the interactions. In these images the Q-sensor ANS recorders are visible on each wrist of all participants. We recorded bilateral signals so that later analyses would be able to test for any effect of hemispheric laterality.

4 The Machine's Task

This section discusses some predictable advances in speech-processing technology that might be of use in future Perceptual Computing. In particular it proposes ways to overcome some limitations in automatic speech recognition, and suggests some improvements that might help to make speech synthesis more interactive.

In a simplified world, a speech synthesiser just has to make an appropriate sequence of speech-like sounds and the user *(horrible word)* or customer *(even worse)* is expected to understand and perhaps act on the linguistic content of

the speech. In the real world however, most speech synthesisers don't even know whether or not there is a listener present! No normal human would start speaking in such a context[1] (unless speaking to oneself, when a listener is either not needed or present by default). The first task of our sentient synthesiser in an ideal human-computer speech-based interface would therefore be to check whether or not speech might be appropriate at any given moment. Perhaps the only thing worse for a synthesiser than speaking into a void is speaking out when silence is preferred, thereby interrupting a human conversation or auditory performance.

Most human speakers will also check that their message is getting across. This does not happen with the typical speech synthesiser of today. People are just expected to understand. A careful synthesiser might even check whether the listener, if present, can actually understand the language being used - the machine might be capable of rendering perfect Chinese, for example, but if the listener is not familiar with that language, then any linguistic utterance generated will effectively be meaningless. There are of course many non-linguistic utterances that are common across many pairs of human languages, but few synthesisers are capable of rendering them appropriately.

So the first five checks that our sentient speech synthesiser should make are, in order: (a) is there a human present and a need to speak?; (b) does that person qualify as a listener (i.e., close enough, with working ears, etc.)?; (c) is the person capable of hearing the sound?; (d) is the person attending to the sound?; and finally, (e) is the function of the sound being appreciated or understood? (which is approaching a philosophical distinction but can be approximately estimated from the synchrony of behavioural responses). If all five conditions are satisfied to a certain level of probability, then the higher-level dialogue component of this speech-based interface can start to estimate whether there is rapport reached between the speaker and listener, or whether a repair is necessary, perhaps some rephrasing of the speech at a different level or genre so that satisfactory comprehension can be achieved.

The above are measures of 'engagement'. In the context of a spoken dialogue, engagement is a feature of cognitive attentional states. Clearly the speaker is engaged; this can often be simply measured by a correlation of mouth movement and presence of speech-like sound. We do not have to pay attention to the content of the speech to know that a person is speaking, and by definition therefore, engaged in that speech.

In Fig. 2 it is perhaps the person in the centre who is speaking. How can we know this, or how would a sensitive machine be able to estimate such information? Perhaps from the shape of the hands. Even in this still image, it is apparent that his hand gesture is supporting a spoken utterance. The person might actually be holding a black bottle, though this interpretation is less likely. But how do we estimate or infer the listeners' cognitive engagement or attentional state? Simply being physically present is not enough. Some difference between hearing and listening must be inferred.

[1] Broadcasters or actors might be an exception to this general rule.

In Fig. 2, both Shannon and Nick appear to be striking a 'listening' pose. They are facing the speaker and have hands either resting or close to an ear. From this still image, we as humans can process much of the visual information and make inferences from these clues about the attentional states of the three people present. It should not be difficult for a synthesiser (or its sensing component) to do the same. If the image is moving, as in a video, then use can be made of coverbal synchrony [7] as the listeners' heads, and parts of their bodies, should be moving in some way that links to the timing patterns and phrasing of the speech.

Laughter can also be a clear indicator of engagement and confirms (if appropriately timed) that the listener is probably paying attention. Nodding, co-gesturing, offering backchannel utterances, interrupting appropriately, etc., can all indicate some degree of engagement in a conversation. So the synthesiser needs to have eyes and ears as well as a 'mouth', but a clever speech recogniser will also be able to make use of these multimodal cues to infer meaning when the actual sound may be too ambiguous to translate[2].

Life will not be easy for our sentient synthesiser; particularly as there may be more than one person present on the scene, and in that case the speaker (in this case a machine) may have to compete with other participants for the right to speak. Some awareness of the cognitive states of the participants will be a necessary part of that dialogue process.

5 The Corpus Annotation Goal

The first use to which we are putting the D-ANS Corpus is for the development of advanced dialogue interface technology. The Q-sensor-derived biometric measures, even at a simple glance, confirm that they can provide useful information for confirming the automatic inference of engagement as estimated from audio and video signals and through use of measures of coverbal synchrony.

Figure 4 shows electrodermal activity (EDA) measures for three people (both wrists) for Day-3 of the recordings. The small grey area near the centre represents one five-minute section of that measure which is shown in more detail for one speaker in Fig. 5. The latter figure shows vertical bars representing speech from each participant (blue for the speaker whose EDA plot is shown above). There is a clear relation between onset of speech and an increase in measured activity. Further relations between the timing of the EDA changes and activity of other speakers is currently being explored in a more systematic way.

We employ statistical means to test these correlations and machine-learning to test the degreee to which audio-visual information can be used as a predictor of the ANS responses as indicated by the EDA signal. To better validate these techniques we also need human-generated annotations of events in the discourse, but this is an expensive and time-consuming task.

[2] Think of the various ways of saying the word 'yes' for example, and the wide range of different meanings they represent!

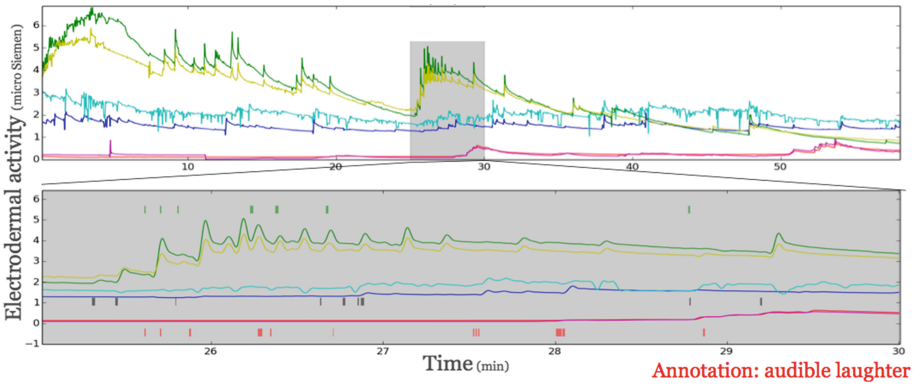

Fig. 4. Data from Day-3 showing electrodermal activity from both wrists of the three participants. The small grey box at the centre-top marks a five-minute section that is shown in more detail in the following figure

Since laughter is such a common feature of casual speech, it is also the first feature that we annotate. The text of each utterance, however, is of lesser importance. It is not really necessary to know the full details of the linguistic content when it is the functional effect of each utterance that most interests us. The social dynamics of a conversation can be equivalent whether the topic of discussion is 'pasta' or 'car engines'; it is the dynamics of turn-taking, and the group involvement that is of most interest to us here, and a simple voice-activity detector (VAD) in conjunction with image processing can provide almost as much information as a full manual transcription in this case. The nature of turn-taking and the length of each utterance can be easily calculated from VAD data which is both visually appealing and machine-friendly for processing. It is also helpful to privacy not to have to reveal too many details of the actual conversations when discoursal dynamics, or conversational metadata, may be sufficient.

The correlations seen in Fig. 5 can be readily detected by automatic processing. The value or meaning of these correlations, however, is what we most need to determine at the present time, and that can only be achieved by human intervention. Our human annotators can determine the tone of the laughter much better than any automatic detector is yet able to; positive supporting laughter, or humorous outbursts contrast with embarrassed or hesitant laughter which might indicate a social negative state.

6 Sharing

The corpus is being made available to interested researchers under the following web-page: http://www.speech-data.jp/nick/d-ans/.

The participants have agreed to share this data with the research community, provided that the details of the personal stories and identifying information

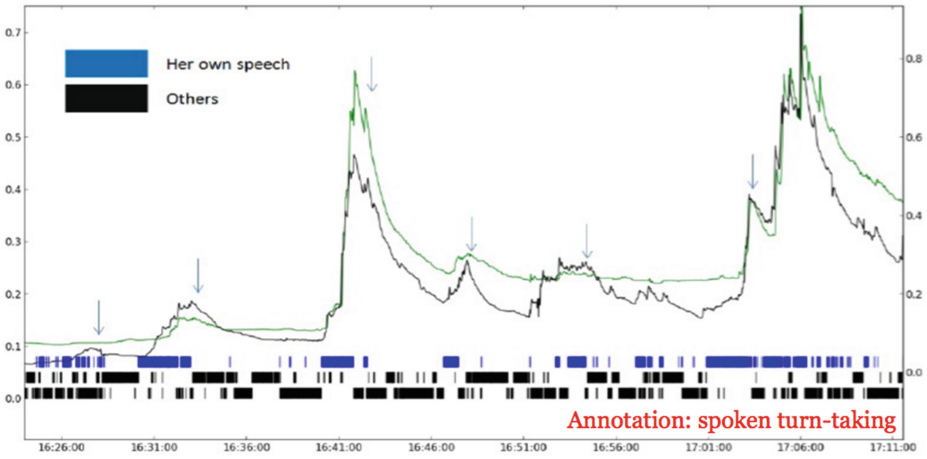

Fig. 5. Five minutes of EDA activity, aligned with speech activity, showing clear spikes coinciding with onset of her speech and turn-taking

(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.

As this is currently active work in progress, the state of the pages is liable to change at short notice, but we invite collaborative study of this material and offer it under a Creative Commons Attribution-NonCommercial International license. The annotations, media, and biosignal data will be shared on the website 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.

7 Summary and Conclusion

This paper is an updated version of an invited talk presented by the first author at the MA3HMI international workshop in Singapore, which brought together researchers working on the analysis of multimodal recordings as a means to develop systems that can interact with humans. The core of the oral presentation was to describe the corpus as originally presented at LREC 2014 [1]. The present paper gives more of the background to the development of the corpus and of the intended uses to which it will be put in our work at the Speech Communication Lab in Dublin and at NAIST in Japan. I am grateful to the organisers of the workshop for giving me the opportunity to discuss these ideas and to Shannon for joining me in the written version of the paper.

We welcome interest in the corpus and are happy to share it within the research community. Collaborative work opens up greater opportunities for further research and the technology is still at a pre-competitive stage where most

benefit can be gained through a sharing of the tasks. It can be commercialised at a later stage when greater understanding of the potentials and limitations of each approach has been achieved.

Acknowledgements. 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). The corpus collection was conducted as part of Shannon's doctoral work, which was funded by Universita degli Studi di Genova and the Instituto Italiano di Tecnologia. The work was co-funded as part of the Japanese Government KAKEN research into MOSAIC: "Models of Spontaneous and Interactive Communication" We are grateful to Fred Cummins and Brian Vaughan and thankful for the annotation efforts of Emer Gilmartin and Celine De Looze.

References

1. Hennig, S., Chellali, R., Campbell, N.: The D-ANS corpus: the Dublin-Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech. In: Proceedings of the ELRA, the 9th Edition of the Language Resources and Evaluation Conference. Reykjavik, Iceland, pp. 26–31 (2014)
2. Beukelman, D.R.: There are some things you just can't say with your right hand. Augmented and Assistive Communication (1989)
3. Pfeifer, R., Bongard, J., Grand, S.: How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press, Cambridge (2007)
4. Affectiva Q-sensors. http://qsensor-support.affectiva.com
5. Dawson, M.E., Schell, A.M., Dilion, D.L.: The electrodermal system. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (eds.) Handbook of Psychophysiology, pp. 159–181. Cambridge University Press, New York (2007)
6. Calvo, R.A., D'Mello, S.: Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans. Affect. Comput. **1**(1), 18–37 (2010)
7. Rojc, M., Campbell, N.: Coverbal Synchrony in Human-Machine Interaction. CRC Press, Boca Raton (2013)