

Machine Processing of Dialogue States; Speculations on Conversational Entropy

Nick Campbell^(✉)

Speech Communication Lab, ADAPT Centre, Trinity College Dublin,
Dublin, Ireland
`nick@tcd.ie`

Abstract. This keynote talk presents some ideas about ‘conversational’ speaking machines, illustrated with examples from the Herme dialogues. Herme was a small device that initiated conversations with passers-by in the Science Gallery at Trinity College in Dublin and managed to engage the majority in short conversations lasting approximately three minutes. No speech recognition was employed. Experience from that data collection and analyses of human-human conversational interactions has led us to consider a theory of Conversational Entropy wherein tight couplings become looser through time as topics decay and are refreshed by speaker changes and conversational restarts. Laughter is a particular cue to this decay mechanism and might prove to be sufficient information for machines to intrude into human conversations without causing offence.

Keywords: Interactive speech synthesis · Human-machine-interaction · Conversational engagement · Laughter · Interactional entropy · Intrusive machines

1 Introduction

People talk with machines a lot. Sometimes intentionally, sometimes unknowingly, and sometimes just for the sheer fun of it. We sit in from of our computers and many of us do actually say things in the direction of that machine, though not always with the intention of being understood. Sometimes those words are unrepeatable. Speech-based interfaces are now ubiquitous. Siri, Cortana, Hi-Google, and the rest, have become tools in our pockets that provide a short-cut to the internet; saving our thumbs for better uses than typing.

In the Speech Communication Lab (SCL) in Dublin, part of the School of Computer Science and Statistics (SCSS) at Trinity College (TCD, the University of Dublin), we are designing machines that know how to talk back. Speech synthesis is an old technology now, and can be found in many places - often unrecognised for what it is - but it is a rare synthesiser that knows it is being listened to. Yet how many people can talk to someone without checking that what they say is being heard, comprehended, understood, taken in?

Our SCL research task is the delivery of information derived from electronic content in various forms, and our need is to be sure that the person we are talking to (‘we’ is a machine in this case) has got the message.

So we use cameras, movement detection, and tone-of-voice changes; any data that we can sense from the outside world to inform our computers that ...
 ...well, in the first place, that there is a person present, and that the person has functioning ears, and is listening ...and that he or she can follow what is being said (i.e., that we speak the same language), and that they are following mentally ...and even (one day) that they have understood.

Then we adapt what we have to say to the way that they are taking it in - perhaps by speaking faster or slower, or by using simpler words, or more intricate ones - adjusting our style of speech and manner in a way that makes it easier for the person to follow. That is the goal. This is still work in progress.

As a first step to learning more about how we should be doing this, we implemented a conversational robot, called Herme, and left it out in a public space to talk with people for three months. That was some years ago.

2 The Herme Dialogues

Herme didn't listen; like many people, she spoke a lot and she watched the person she was talking to, so as to sense their reactions, and then she just carried on speaking - drawing the interlocutor into her dialogue but not paying much attention to their replies. She could keep people 'chatting' with her for about three to five minutes before her conversation came to an end. She asked simple questions like "What's your name?" and "Why are you here today?" and waited while they replied, sometimes interjecting a "Really!" or an "Oh?" to keep them talking. Her main task was to get them, eventually, to sign a consent form so that we could use the material we were filming of their interaction. By showing them that she had found their face in her environment¹, she managed to persuade people that she was listening to them. She was certainly watching them ...she needed to see when their face stopped moving so that she could start her next utterance. A very simple technique, but one that we found most effective.

We collected dialogues from about 1500 people of whom about two-thirds voluntarily signed our consent-form. All were recorded. Laughter was common. People were charmed by her voice (like that of a small child) and she was cute, and told them a joke, and even managed to get them to tell her a joke themselves; well, a 'knock-knock' joke anyway. Anyone can tell a knock-knock joke. And everyone laughs when they're chatting. Herme has a cute laugh.

Laughter seems to be a special form of lubricant that keeps the conversations going, but not all laughter has to do with jokes. People laugh when they're embarrassed, when they don't know what to say (if they are relaxed) and when they get the point of what you are trying to say to them. Laughs work as a sign that the conversation is going well. They're a great signal to process.

¹ There was a large screen behind Herme's stand showing passers-by what she could see, with a coloured circle drawn around each face in the scene.

3 Conversational Speech Synthesis

Herme used a very old speech synthesiser, Apple's '*Princess*' voice, warped by compressing the formants and raising the pitch to make it sound as if it came from a smaller body. We used a hardware filter for this² but it is trivial to do in software nowadays.

For laughs, she could only say "tee hee hee" and "ho-ho-ho" but it was enough. People responded to her laugh with great warmth and it relaxed them enough to keep them listening through the next stage of her spiel.

Herme was a testbed for one type of conversational interaction, but the need for more flexible conversational speech synthesis is probably great and growing. Machines must learn how to speak. They can talk already; talking machines have been around for a long time, but speaking is different: speaking needs a partner. A partner is not the same as a listener. Students listen when the professor speaks, but that is a complicated form of partnership. Most people speak in informal environments, and they intersperse their speech with chat.

We have shown that a machine can chat with a person - Herme was proof of that - but it was an unbalanced conversation. The robot took the lead and the conversation didn't get beyond the early getting-to-know-you stages. She couldn't have held a sustained conversation or spoken easily with the same people on different days without being caught out.

The machines that we use to deliver our spoken digital content will need to have a memory of what has been said (or spoken about) and will need a sense of timing or knowing when to speak. There are strong social rules for that.

3.1 A Talking Fridge

Let's imagine the smart home of the near future. Each room is wired up with sensors that stream information into the home-server (a computer that links the home with the outside world of information). It will probably be part of the fridge. The refrigerator is the one device that doesn't get switched off when people go out so there'll be a constant supply of power. The fridge is also in the place where people gather most. They might relax in front of a large screen but they probably eat round a table in the kitchen. It's the family place.

Like Herme, the fridge (or the home-server, an interactional device) can monitor what is going on in the room around it. It doesn't have to listen to what is being said or talked about; just know enough about who is doing what to be able to interrupt with a message if it has one. If the people around the table are deep in conversation (i.e., their heads are moving in a certain pattern and sounds are being made), then it might be wiser to wait for a lull in the talk before butting in with what it has to say. If they are watching the news, then it might be better to wait until the adverts come on.

It would good for our machine if it had a notion of conversational states and of the types of engagement of each conversational participant in the real world around it. The fridge needs awareness.

² A Roland Sound Canvas.

3.2 Entropy (An Interlude)

Erwin Schrodinger [5] was at TCD when he gave his lectures on “What is life? The Physical Aspect of the Living Cell”³. He said:

Every process, event, happening - call it what you will; in a word, everything that is going on in Nature means an increase of the entropy of the part of the world where it is going on. Thus a living organism continually increases its entropy - or, as you may say, produces positive entropy - and thus tends to approach the dangerous state of maximum entropy, which is of death. It can only keep aloof from it, i.e. alive, by continually drawing from its environment negative entropy -which is something very positive as we shall immediately see. What an organism feeds upon is negative entropy. Or, to put it less paradoxically, the essential thing in metabolism is that the organism succeeds in freeing itself from all the entropy it cannot help producing while alive.

Conversation is a living organism. Entropy kills conversation. Laughter reduces entropy by resetting the topic, and so keeps conversation alive.

He also said (in the same lectures):

The disintegration of a single radioactive atom is observable (it emits a projectile which causes a visible scintillation on a fluorescent screen). But if you are given a single radioactive atom, its probable lifetime is much less certain than that of a healthy sparrow. Indeed, nothing more can be said about it than this: as long as it lives (and that may be for thousands of years) the chance of its blowing up within the next second, whether large or small, remains the same.

but this is a matter for discussion elsewhere.

Google’s definition of entropy :

entropy

/ˈɛntərpi/ 

noun

1. **PHYSICS**
a thermodynamic quantity representing the unavailability of a system's thermal energy for conversion into mechanical work, often interpreted as the degree of disorder or randomness in the system.
"the second law of thermodynamics says that entropy always increases with time"
2. lack of order or predictability; gradual decline into disorder.
"a marketplace where entropy reigns supreme"

³ Lectures delivered under the auspices of the Dublin Institute for Advanced Studies at Trinity College, Dublin, in February 1943.

4 A Notion of Conversational Entropy

After Herme, we became more interested in laughter and especially how it punctuates a discourse. Francesca Bonin’s PhD [2] examined the structure of conversational interaction and explored the relation between social signals and discourse phenomena such as topic changes, investigating whether social signals have a discourse function in addition to their social function. Different analyses that investigated the temporal dynamics of laughter, backchannels, silences and overlaps, were explored, finding a relation between topic changes and a decrease of social signals. Specifically, it was found that immediately after a topic change there is a significant drop in social activity, defined by her as interactional entropy: “*The interactional entropy of a segment x is defined as the number of occurrences of social signals in x* ” (ibid p.71).

Through comparing topic changes in two corpora of spontaneous spoken interaction, she concluded that a constant trend emerges in both TableTalk and AMI: topic terminations (*wt*) show a significantly higher presence of signals if compared to topic beginnings (*wb*). In AMI [4], among all the distributions of frequencies of laughter, overlaps, silences, lexical and non-lexical backchannels in *wt* and *wb* the non-parametric Wilcoxon test rejects the null hypothesis of $wb = wt$ and validates the alternative hypothesis of $wb < wt, p < 0.0005$. In TableTalk [1] the same applies to laughter, overlaps, and lexical backchannels. In other words, topic terminations reveal higher interactional entropy than topic beginnings.

In fact a drop in social signals appears to occur immediately after a topic change when the interactional entropy [...] is reduced. Participants show the tendency to limit the interaction immediately after a topic change, probably to leave the floor to the speaker who has introduced the new topic (ibid p.128).

She clearly showed that after a topic change a decrease of interactional entropy occurs, and concludes that this information might be used to better understand the discourse structure via non-linguistic information such as laughter, overlaps, backchannels, and silence, and thereby shed new light upon the discourse functionality of social signals.

It seems that introducing a new topic reduces the entropy of a conversation (‘feeding it’, as Schrodinger would say). Conversely, by observing the amount of non-verbal behaviour in speech (particularly laughter) we can estimate the likelihood of a forthcoming topic change, and thereby enable our device to interrupt at a timely point without having to listen in on the actual content of any conversations.⁴

The system can be aware of its environment through sensing movement and the dynamics of vocal activity around it. It doesn’t need to listen. Perhaps that is what many people do too? Conversation is a uniquely human form of behaviour.

⁴ The idea that household devices might be capable of eavesdropping on nearby conversations is rightly anathema to many kitchen owners and occupants.

5 Social Interactions and Signal Processing

For our system though, a conversation is a data source; a signal that is available to be processed. With Herme, we avoided the use of ASR (automatic speech recognition) for several reasons; it often fails in a noisy environment, it needs specialised domain dictionaries and language models, and it is intrusive. It is the last point that is of most concern to us now. Herme was in a public space and engaged in trivial social chat with a large number of unknown people. Nothing sensitive or really personal was discussed. In the home though, the situation is different. The potential for misuse of available information has been much in the news recently and people in general are now becoming quite wary of devices that leak or pass on information. The law may be clear (voluntary sign-in usually absolves the supplier of legal responsibility), and the ethical issues are certainly of concern to most scientists, but the technology must be made watertight against leaks if our work is to be trusted in society-at-large.

5.1 Natural Human-Machine Conversational Interaction

The ‘listening & watching’ fridge that may host our technology in the future should be able to observe the goings-on in its environment much as a pet dog may watch and be aware of the happenings of the home. It will of course have to ‘listen’ carefully when commands or instructions are given, but when in ‘sleep mode’, it should not be hearing everything.

The work presented above may offer a solution to this conundrum. If the device keeps a measure of the entropy of conversations in the home, without listening to what is being said, through processing of non-verbal and behavioural information, then it can perhaps be considered safe.

At the same time, the amount of processing that is required from a ‘conversational agent’ can also perhaps be significantly reduced; if the machine only has to devote energy to processing linguistic/semantic propositional information at certain isolated points in the signal then its energy can be greatly preserved, and more time may be devoted to the arduous symbolic processing needed to ‘understand’ speech.

By maintaining an awareness of the social energy in its environment, perhaps our speaking device will appear well-mannered, only interrupting when necessary and maybe often with a delicate or appropriate sense of timing? It might be far-fetched to imagine the machine joining in with a joke as Herme did, but if it has the sense to ‘understand’ what processes are happening in the human sphere, then like a pet cat or dog at home, it might be a welcome guest.

6 Conclusions

This paper describes some ideas to be presented in a Keynote at Specom 2016. The invitation tentatively specified “Overview of speech technology results, challenges, trends, promising directions in Social Interactions and Signal Processing”

as a title. We chose instead to present some current work from our lab in Dublin as the basis for speculation about higher matters. The facts of current research are perhaps well represented by other papers in these proceedings.

The concept of entropy was introduced at the beginning of the previous century and has been well-understood by physicists, chemists, and information engineers, among others, but has failed to take hold in the humanities. This is sad. Our entire world is subject to entropy, and its concepts may throw light on more than mere mechanics or thermodynamics. The actions of people in society, and particularly the structured actions of participants in conversation are subject to the same laws, and the same probabilistic processes.

Addendum Gibbs' definition of free energy : (something good to think about)

the greatest amount of mechanical work which can be obtained from a given quantity of a certain substance in a given initial state, without increasing its total volume or allowing heat to pass to or from external bodies, except such as at the close of the processes are left in their initial condition [3].

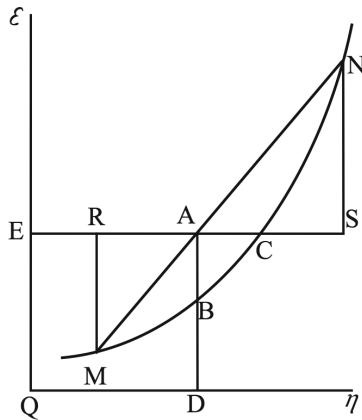


Fig. 1. Graphical representation of the free energy of a body. The figure shows a plane of constant volume, passing through the point A that represents the body's initial state. The curve MN is the section of the “surface of dissipated energy”. AD and AE are, respectively, the energy (ϵ) and entropy (η) of the initial state. AB is the “available energy” (now called the Helmholtz free energy) and AC the “capacity for entropy” (i.e., the amount by which the entropy can be increased without changing the energy or volume). From Gibbs, J.W. (1873). “A method of geometrical representation of the thermodynamic properties of substances by means of surfaces”. Transactions of the Connecticut Academy of Arts and Sciences 2: 382–404., Public Domain, <https://commons.wikimedia.org/w/index.php?curid=3279793>

Acknowledgments. This research is supported by Science Foundation Ireland under Grant No. 13/RC/2016, through the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Trinity College, Dublin. We are grateful to the School of Computer Science and Statistics at Trinity College Dublin for their support of the Speech Communication Lab.

References

1. Tabletalk is a multimodal multimedia corpus of free flowing natural conversations, recorded at the advanced telecommunication research labs in Japan (2005). <http://sspnet.eu/2010/02/freetalk/>
2. Bonin, F.: Unpublished Ph.D. thesis: “Content and Context in Conversations: The Role of Social and Situational Signals in Conversation Structure”. Trinity College Dublin, Ireland (2015)
3. Gibbs, J.W.: A method of geometrical representation of the thermodynamic properties of substances by means of surfaces. Connecticut Academy of Arts and Sciences (1873)
4. McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al.: The AMI meeting corpus. In: Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, vol. 88 (2005)
5. Schrödinger, E.: What Is Life? the physical aspect of the living cell and mind. Dublin (1943)

Speech and Computer

18th International Conference, SPECOM 2016,

Budapest, Hungary, August 23-27, 2016, Proceedings

Ronzhin, A.; Potapova, R.; Németh, G. (Eds.)

2016, XVIII, 731 p. 197 illus., Softcover

ISBN: 978-3-319-43957-0