Herme & Beyond; the Collection of Natural Speech Data

Nick Campbell
Speech Communication Lab
School of Computer Science & Statistics
Trinity College Dublin, Ireland
nick@tcd.ie

Abstract

This paper describes our approach to the collection of 'natural' (i.e., representative) data from spoken interactions in a social setting in the context of the development (through time) of expressive speech synthesis. Over the past ten years or so, we have collected several corpora of unprompted social conversations that illustrate the 'contact' element of speech that was lacking in many of the corpora collected by use of a specific 'task' with paid participants. The paper discusses the technical and ethical issues of collecting such spoken material, and highlights some of the problems we have encountered in the processing of this much-needed data. Through the use of attractive conversational devices, we have found that natural human curiosity, and an element of social programming combine to provide us with a rich source of material that complements the task-based collections from paid informants.

Keywords: task-free dialogue, spontaneous speech, data collection techniques, ethics & incentives, the common good

1. Introduction

Developing speech synthesis technology requires the collection, annotation, and analysis of large amounts of speech data and as our knowledge of speech processes grows, alongside a phenomenal growth in computer memory, processing power, and bandwidth, we find an ever-increasing need for larger amounts of material. Gunnar Fant, perhaps the founder of computer speech synthesis, firmly understood the science of voice production and built his talking machine from first principles, with no need for a corpus of examples to replicate. Denis Klatt, in his seminal work of the 80's studied copious spectrogram printouts of actual vocalisations to increase the naturalness of his speech output by modelling the features and dynamics that he observed in the data. Joe Olive, another pioneer of this field, used actual recordings from which he cut diphone-sized segments of speech for a more precise modelling of the information carried in the transitions between the phones. (Fant, 1970; Klatt, 1987; Olive 1980)

The 80's saw the development of machine-learning and increased use of statistical modelling with the consequent growth of multinational initiatives for the common collection of speech material from across the world, and the development of organisations such as the LDC and COCOSDA (with ELDA/ELRA coming close on their tails) and the recommendation of specifications and tools (BLARKS) for the collection and annotation of common speech material (Krauwer 1998; Mapelli 2003). On this foundation, the present ubiquitous speech technology was built.

The situation in the present century is vastly different; and the foundation technology that grew from common experiments has become integral in mobile devices and ubiquitous human interfaces. Corporations use these technologies for their daily interaction with customers and now stream almost infinite amounts of real-world data through their systems. Deep nets have evolved to process this information on massively parallel gpu devices that make the small collections of the immediate past seem very primitive in comparison. The world of speech processing is now split in two; those that provide actual services can find more data than they need, while those in academia or smaller industrial start-ups are left with no access to the corporate streams. In parallel, ethical issues of data collection, storage, and protection arise to a frightening degree, as the potential for abuse (or leakage to unintended recipients of confidential information) becomes a more real everyday threat. We must now find new ways to collect corpora (or learn from live streaming speech processes) that meet modern size requirements yet preserve privacy.

How did speech data collection become a threatening activity? What happened in the transition from innocent spectrogram analysis to privacy-revealing spontaneous speech collections? In the pre-Snowden age, recordings were not treated with suspicion or fear. Subjects gladly contributed; as 'giving your voice to science' was on a par with 'leaving your body for medicine' and was considered an altruistic act, not necessarily requiring payment. Some incentives were provided (Call Home collections for example (Canavan 1997) offered cut-rate or free calls) but the amounts of data were relatively small and the content, even if of a personal nature, was considered privileged and not open to abuse.

2. Expressive Speech Processing

Yoshinori Sagisaka of ATR in Japan introduced the ν-talk system of non-uniform concatenative speech synthesis (Sagisaka et al 1992) based on recordings of 5000 words and 503 sentences as raw material. This was considered a large corpus at the time. The recordings were from professional announcers, people trained to produce consistent 'standard' pronunciations in a 'received' quality of voice. There were no hesitations in the readings, and no laughter or other non-speech vocalisations. The recordings were segmented by hand and strict labelling applied; the phoneme set was known, and allophonic variation was taken care of automatically as being due to phonetic context dependencies. The resulting synthesis was clear, well-

articulated, and pure-'Tokyo'! No regional or personal deviations were allowed.

These methods of speech synthesis produced clear formal-sounding sentences (each utterance had a well-marked full-stop at the end!) suitable for announcements, broadcast-news reading, and impersonal information provision. They sounded robotic because a) the signal was manipulated, and b) the text was 'unnatural'. But that was the nature of speech synthesis at the time. These were Talking Machines that rendered text into speech, based on the dream of reading machines from the 70s. There was no need for laughter or hesitations as these were perceived as speech 'defects' and the sign of an untrained speaker or an amateur performer.

With the growth in the availability of speech recordings we were able to extend ν-talk to produce CHATR and by removing the signal processing to instead use raw speech segments in unprocessed form for concatenation were able to reproduce the known voice of any given speaker (see the paper on 'CHATR the Corpus' in the main conference). This brought with it the sometimes embarrassing facts of 'natural' speech that varied from the 'received' dialect/accent and displayed all manner of 'spontaneous' speech phenomena as were found in the original recordings. Talking Machines had become capable of conversational speech.

Given that the synthesiser was now able to replicate any voice, dialect, or speaking style, the question remained as to what types (variations beyond the mean) would be required for conversational speech synthesis. Even for reading books, a considerable range of voice qualities and expressivity would be required; but for 'interactive' synthesis where the machine would need to replicate human characteristics of speech, the territory was uncharted. Would the machine need to laugh, for example? Would it need to cough? Singing synthesis was already being explored elsewhere as an independent field of research, and poetry-reading was perhaps too specialised a form of vocalisation to require synthesis. The limits were unknown and hence the need for representative corpora.

The conundrum here was that a well-designed corpus would produce all the examples that it was conceived to collect, but there was no specification of what that coverage might require. On the other hand, an undesigned corpus was at that time a contradiction in terms, leaving too much to chance. We now have task-specified collections where applications stream in data from countless users, but when the base technology was still under development, that was too ambitious to even consider. The Table-Talk Corpus was a first attempt at resolving this data collection problem.

2.1. Table Talk

Table-Talk (ISLRN: 545-953-122-584-3) was an early experiment in multimodal speech data collection. Five participants met over a period of three days to sit together and talk surrounded by microphones and cameras that recorded everything from several angles. No task was specified and no topic set. Here we discovered the wonderful facility that humans have for just talking (Dunbar 1998). Silence in social situations is taboo, so people sharing a common space start spontaneously to chat. No new science was involved but the data we collected showed intriguing patterns of interaction dynamics and vocal usage. There were vey few 'well formed sentences' among these utterances. Instead there was a rich variety of laughter and spontaneous 'chirping' as topics emerged and interest grew around them. Topics decayed away to be replaced by others, arising from points previously raised, or completely introducing a new subject.

This experience emboldened us to propose the Expressive Speech Collection (funded by the JST) whereby people volunteered their speech in exchange for token payments (and the possibility to keep the recording device (a mini-disk recorder) for personal use). No constraints were made on the speech to be recorded and participants were encouraged to keep the recorder active at all times so that when an interesting event occurred there would be no need to interrupt the flow by switching on the machine. Although there are strong ethical constraints on deliberately inducing fear in participants, the recording of natural fear (in the case of an earthquake for example) was considered inoffensive. In the five years we were recording there was not one fear-inducing quake, but several minor tremors. Similarly 'joy' and 'surprise' can be difficult to elicit (fake?) in the studio but are common occurrences in nature. We had faith that what was being collected would be representative of the types of vocal activity that would be needed by a speaking machine that was to operate in the real world, perhaps taking the part of a remote human in a local (and possibly translated) conversation.

2.2. JST ESP

The findings of the JST/ESP data collection (Campbell 2002) have been reported widely elsewhere. Sufficient here to note that they revealed a wealth of unexpected facts about how the voice is used in social situations in the real world. They also revealed the extent to which non-verbal information is used in place of linguistic structures, and how the social element in interaction absolutely dominates for most of the time. There were few extremes of joy or sadness but plenty of everyday expressive speech and many meaningful variations in voice-quality and speaking style. Previous ideas of how spoken interaction worked had been based on linguistic components alone and a new field of expressive interaction was opened up. Previous recordings of spoken interaction had been predominantly task-based, and the participants (being paid for their expensive time) were usually loath to digress from the specified task to 'waste time' in 'mere' social chit-chat! In this context it is interesting to note the difference between Petukhova's PhD thesis (Petukhova and Bunt 2012) and the resulting ISO standard that arose from it ISO 24617-2 makes no reference to 'contact events', whereas two levels of interaction in the thesis depend on them. The ISO standard lacked evidence for social contact because the majority of the corpora that had been collected were planned top-down and specified the tasks (and therefore the coverage) of the speech in advance. No social contact occurred. The paradigm itself renders the collected speech unnatural in a social sense.

2.3. D64

From the ESP insights we gained on the value of spontaneous interaction it was a short step to the recording of the D64 Corpus in Dublin (Oertel 2010). We booked a hotel apartment for three days (number D64) and populated it with equipment and people. No money changed hands, and no instructions were given, though each participant did sign a consent form acknowledging that everything was to be recorded, warning that indiscretions were inadvisable, and giving each the right to withdraw at any time or have recordings erased from the record if so desired. Food and drink were provided (including wine on the third day!) and devices were left running from before the start to after the end (the setup and calibration of various recorders actually makes a particularly interesting part of the corpus as stresses were high given the time constraints and technical complexity of the equipment). The participants quickly became friends, sharing some extremely personal information at times, and no thought was given to forms of payment - this was fun! But the participants were all academics - and there is a general expectation in this community that effort is to be freely contributed (paper reviews for example) towards the greater good of generating knowledge

2.4. D-ANS

Perhaps this philosophy underlay the Dublin-Autonomous Nervous System Corpus of Biosignal and Multimodal Recordings of Conversational Speech (D-ANS: Hennig 2014), as the participants were members of the same lab, taking a break and chatting in front of cameras while wearing biosensors. The conversations that arose were without doubt 'natural' and completely spontaneous, and the biometric readings that we collected in addition to the audio and video data again revealed patterns of the cognitive processes underlying social speech production that were not known beforehand. This was not 'work' per se but a voluntary effort on a very small scale to increase our understanding of speech processes. The challenge now, having learnt the worth of spontaneous and informal collections is to generalise them to a larger scale and to automate the subsequent processing. A manually segmented and annotated corpus of even this small size can take several years before coming to fruition (Gilmartin et al 2013).

3. Herme & beyond

Herme was different. Here we employed one-to-one conversations instead of group talk, and we had no idea how many participants would take part (Han et al 2012).

Herme was a small motorised ©LEGO robot platform that supported a web-cam (with high quality microphone) and triggered a new conversation when a person was spotted (by use of OpenCV face recognition). The device was exhibited as part of a three-month exhibition (Human+) in the Science Gallery in Dublin; a high-tech art space where members of the public can come in from the rain to enjoy science & technology with some coffee and free wifi.

We maintained total control of the conversational flow from the start, as Herme always took the initiative, listened to any responses (without ASR) and responded with a backchannel or changed the subject according to a predetermined sequence of conversational utterances. Both wOz and automated versions were tested but the human operator proved significantly better than the algorithm at keeping a participant interested in the conversation. The sequence of utterances was identical in both paradigms but the timing of utterance onset was too delicate a control for the software to compete. While not the focus of the current paper this aspect of timing control for dialogue speech synthesis is currently work in progress, and the data from 'failed' conversations is invaluable for training statistical models.

Natural curiosity was probably the main incentive driving most of the Herme coversations - people were attracted by the object - it moved, made noises and most importantly had a display which showed what it saw. When a person approached, their own face appeared in Herme's display, with a circle drawn round it to show that she[1] had recognised them as a person. When Herme spoke at that point it was not immediately natural for people (as observers) to respond, but when she repeated the greeting most people responded with a greeting in return - accepting on the second utterance that the robot was talking to them and becoming active participants.

The voice of the robot was childlike, and the childlike innocence (and directness) of the questions she asked had an appeal that many people instinctively responded to - and answered politely (or jokingly) in response. Most participants stayed for about three minutes, the length of a complete conversation, and then signed a release form giving researchers permission to use the data when asked to do so by the robot. It was clearly signposted that all conversations were being recorded. Over the three-month period more than 1500 people voluntarily took part in a conversation with the robot and about two-thirds signed the consent forms to allow us use of their data.

Gilmartin & Su (forthcoming) have recently extended Herme to produce *'Cara'*, a conversational autonomous relational agent, which was recently exhibited[2] as part of the All Ireland Linguistic Olympiad at Trinity College in Dublin. This software instantiates a full dialogue system and uses ASR in conjunction with Voice-Activity Detection to inform the dialogue manager of which utterance to render next and at what time. The Olympiad attracts some of the brightest and most inquisitive of Irish schoolchildren to compete on linguistic puzzles and our side-exhibition provided a rich source of interaction behaviour as the children took turns to chat with the robot during their breaks.

The experience was mutually beneficial - the curiosity of the children prompted them to test the limits of the robot's dialogue capabilities, providing a learning experience for both sides, and fun for the participants while producing invaluable data for the developers. Of course the system failed often - the state of the art in autonomous dialogue systems is still far from ideal, but from the point of view of research, if everything runs smoothly then there is little left to learn, and as our goal in collecting these data is to gain experience, then failure (of a dialogue) is as valuable to us as 'success'.

[1] Herme is generally thought of as 'female'

[2] mid-March 2016

4. Discussion; Generalising the Process

A complaint from an industry representative at a recent Interspeech lunch was that many of the scientific papers were reporting results from corpora of less than 20-hours of speech material, pointing out that results from such small studies just don't generalise to be useful for solving real-world problems. He might have said 200-hours, the point would have been the same.

Corporate analysis of speech data reported at a recent ICASSP cited 200,000 hours of speech material as normal for training. The major service providers have solved the data collection problem and are now tackling the issues of working with really 'big'-data but are unable for a variety of reasons to make that resource available to a wider public. Nor do they perhaps see the need to solve some of the problems that academic researchers find interesting.

Fortunately many corporations take in interns for short periods and experiments can be made (under strict limitations of confidentiality) on in-house data ("this call may be recorded for training purposes") the general results of which can be published more widely.

Social media also provide rich streams of interesting material but apart from the technical and legal problems with tapping these sources, the 'language' they use is perhaps unique to the medium. It may be evolving to form a common subset of human language with its own grammar and syntax (hash-tags, etc.,) but is less useful for synthesis.

The need for task-based conversational data can presumably be satisfied by the applications that provide the services that meet the tasks, but there is still a need for non-task-based, primarily social speech data for the next generation of human-machine interfaces. Machines may not need to replicate the full range of human sounds in synthesised speech but they will, we argue, be required to process this information to make inferences about the human cognitive states in an interaction so that an appropriate response may be served by the machine.

5. Conclusion

This paper has described our approach to the collection and analysis of speech data for the development of interactive speech synthesis for use in dialogue systems. We firmly believe that it is of more value to collect unstructured data that yields fresh knowledge on speech processes and that the top-down design constraints of a 'well-designed' corpus can prevent these spontaneous natural features from emerging. As our systems develop, so we can use them to collect more material. The element of fun in interacting with a machine in a very human way seems to motivate people to help us, and we learn much from what they try to make the machine do. The types of voice, speaking-style, and vocal activity have surprised us in the ways they deviate from standard descriptions of linguistic use. We infer that the linguistic models, and the types of speech that synthesisers are generally trained on are abstracted away from the complex details of everyday performance and encapsulate instead a higher knowledge about the language and speech per se, rather than an encoding of actual everyday performance. The value of collecting data in the wild far exceeds any financial or other costs and will, we hope, help us to provide an interface that is more in touch with the actual everyday needs and expectations of the people who will have to use this technology in speaking devices of the future.

Acknowledgements

This work has been carried out at ATR in Japan and in the Speech Communication Lab in Dublin with funding support from JST, Kaken, SFI, CNGL, and the ADAPT Centre. The work owes much to the many inspirational researchers who have contributed time and effort and helped in this great learning experience.

References

Fant, G., (1970) Acoustic Theory of Speech Production. Mouton De Gruyter. ISBN 90-279-1600-4

Klatt, D., (1987) "Review of text-to-speech conversion for English" J. Acous. Soc. Amer. 82, 737-793

Olive, J., (19980) "A scheme for concatenating units for speech synthesis", in Proc Acoustics, Speech, and Signal Processing, IEEE International Conference ICASSP '80. (Volume:5) Apr 1980, pp.568 - 571

Krauwer, S., (1998) "ELSNET and ELRA: A common past and a common future", in ELRA Newsletter Vol. 3 N. 2.

Mapelli, V., Choukri, K., "Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps", ENABLER project internal report, Deliverable 5.1, 2003.

Sagisaka, Y., Kaiki, N., Iwahashi, N., Mimura, K. (1992) "ATR nu-talk speech synthesis system". Proceedings, International Conference on Spoken Language Processing.

Canavan, A., David G., and Zipperlen, G. (1997) "CALLHOME American English Speech" LDC97S42. DVD. Philadelphia: Linguistic Data Consortium.

Petukhova, V. and H. Bunt (2012) The coding and annotation of multimodal dialogue acts. In Proceedings 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul.

Oertel, C., Cummins, F., Edlund, J., Wagner, P., and Campbell, N. (2010) "D64: A corpus of richly recorded conversational interaction". Journal on Multimodal User Interfaces, pages 1 – 10.

Dunbar, R. (1998). Grooming, gossip, and the evolution of language. Harvard Univ Press.

Campbell, N., (2002) "The Recording of Emotional speech; JST/CREST database research", in Proc Language Resources and Evaluation Conference (LREC).

Gilmartin, E., Hennig, S., Chellali, R., and Campbell, N. (2013). Exploring sounded and silent laughter in multiparty social interaction - audio, video and biometric signals. Valetta, Malta, October.

Hennig, S., Chellali, R., and Campbell, N. (2014). The D-ANS corpus: the Dublin-Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech. Reykjavik, Iceland.

Han, J.G. et al. (2012) "Speech & Multimodal Resources: the Herme Database of Spontaneous Multimodal Human-Robot Dialogues". 8th LREC, Istanbul, Turkey, 23-25 May.