The Edinburgh Speech Production Facility DoubleTalk Corpus

James M Scobbie[1], *Alice Turk*[2], *Christian Geng*[3], *Simon King*[4], *Robin Lickley*[1], *Korin Richmond*[4]

[1] CASL Research Centre, Queen Margaret University, Edinburgh, Scotland, UK
[2] Linguistics and English Language, University of Edinburgh, Scotland, UK
[3] Department Linguistik, Universität Potsdam, Germany
[4] CSTR, University of Edinburgh, Scotland, UK

turk@ling.ed.ac.uk

Abstract

The DoubleTalk articulatory corpus was collected at the Edinburgh Speech Production Facility (ESPF) using two synchronized Carstens AG500 electromagnetic articulometers. The first release of the corpus comprises orthographic transcriptions aligned at phrasal level to EMA and audio data for each of 6 mixed-dialect speaker pairs. It is available from the ESPF online archive. A variety of tasks were used to elicit a wide range of speech styles, including monologue (a modified Comma Gets a Cure and spontaneous story-telling), structured spontaneous dialogue (Map Task and Diapix), a wordlist task, a memory-recall task, and a shadowing task. In this session we will demo the corpus with various examples.

Index Terms: discourse, EMA, spontaneous speech

1. Introduction

We announce the release of the DoubleTalk Corpus, available online for free download at http://espf.ppls.ed.ac.uk/. The first release comprises orthographic transcriptions of phrasal chunks, aligned acoustically to the start and end of the phrase, with speech data from two synchronized Carstens AG500 EMAs (Electromagnetic Articulographs).

Full technical details of the facility are available in [1], but briefly, the EMA machines each record 3D positions and rotations of 12 sensors attached intra-orally and on the head every 5 ms. The machines are mutually synchronized by capturing (a) synch impulses of both machines and (b) the acoustic waveforms of both speakers by means of Articulate Instruments Ltd. hardware. After correcting for TCP-IP inter-machine communication delays, inter-machine asynchrony is better than 1 ms. The hardware is also capable of synchronizing other time series data (such as EPG). The EMA machines are positioned 8.5 m apart to avoid electromagnetic inter-machine interference.

The ESPF facility enables detailed comparison of the nature and time-course of both articulatory and acoustic aspects of each participant's speech, including their interaction as discourse partners. Though previously a single speaker pair was reported [2], we believe this is the first corpus of its type.

Full materials and details are available online at http://espf.ppls.ed.ac.uk/ along with the speech data. Analysis can be performed with any of the many tools developed for Carstens EMA data. However, a special module of the Articulate Assistant Advanced software package [3] was specially designed for this corpus to let it be viewed and analysed via a standard interface (also used for electropalatographic, video and ultrasound data types). Annotations are presented as PRAAT textgrids [4].

2. Speakers

Five of the six speaker pairs comprise naïve non-linguist participants, both native speakers of Standard British English, one with a Southern English accent (SSBE / RP), the other with a Scottish accent (SSE). Neither speaker knew the other. All had previously undergone a process of accent screening and familiarization with EMA as part of the consent-granting process. As well as speaker information, we collected speech data (acoustics only) using the QMU version of Comma Gets a Cure and our wordlist (details below). The final pair (Northern English General American) was recorded as a pilot. The speakers are non-naïve and know each other.

Supplementary information includes a digit-span score of short term memory, an Empathy Quotient score [5], and articulatory data from a short diadochokinetic task.

3. Speech tasks

The corpus includes spontaneous monologue, spontaneous conversation, repetition from memory, shadowing, and read speech. The tasks were designed to exemplify a wider range of styles of speech than usual (e.g. in collections of read sentences), to elicit a greater variety of normal speech production, to add to the diversity of English accent data, and to enable study of articulatory aspects of naturalistic conversation. In particular, the synchronised data from the spontaneous and discourse elements of the corpus are novel.

Since the speakers were in different recordings studies, a flexible audio-only talk-back system was designed to let them interact as required (without visual interaction).

3.1. Monologue / non-interactive tasks

Scripted tasks were included to guarantee some baseline data, and to facilitate comparison to previous EMA datasets. The phoneme-rich story that was used, Comma Gets a Cure [6], had been adapted for Scottish English at QMU. A wordlist was used for the same purpose, organised around lexical sets [7].

A spontaneous monologue was also collected from each participant. This story telling task varied from speaker-to-speaker, and elicited either a personal anecdote, or a retelling of a familiar story such as a fairy tale.

3.2. Dialogue / interactive tasks

Materials for two standard structured tasks were designed, with the main purpose of eliciting spontaneous conversation. The designs were intended to elicit particular lexemes which were judged likely to provoke some conversational difficulties between the speakers, given the mix of Scottish and English accents. Two versions of each task were run.

First, map tasks [8] were used. These alternated so that each participant had one chance to be the information giver. Phonetically ambiguous landmark names, minimal pairs, missing information and contradictory locations was used as conversational foils. For example, a Scottish follower had landmarks such as both *fishing bait* and *fishing boat*, with the English information giver had *fishing boat*. A typical RP fronted GOAT-vowel might cause some misperception or confusion, leading to multiple repetitions. The map path was visualized on the prompt screen via a graphics pad, maintaining the head-up orientation of the follower.

Second, a spot-the-difference picture task based on the Diapix model [9] was used. Each speaker had their own version of a scene, and around ten differences had to be collaboratively found. In addition to some of the same techniques used in the map task to stimulate longer and more interactive conversation, some tongue-twister landmarks were included.

In addition, a short memory-taxing story recall task was used. The information-giving speaker in this task had to read a three-sentence story, one sentence at a time. The follower had to recall each sentence with complete accuracy, and the story-teller was able to repeat and help until the complete prompt could be recalled. Again, the sentences included some lexemes which might cause inter-dialectal difficulties as an extra load. The roles were then reversed for the second iteration.

Finally, a shadowing task was used. This happened simultaneously with, and was based upon, the spontaneous monologue story-telling of the other speaker. They were unable to hear the shadowing, so performed their spontaneous monologue without competition. The shadowing speaker could hear both themself and the monologue, and was instructed to repeat what they heard accurately and soon as possible after hearing it.

4. The Corpus

4.1. Characterization of the speech data

For the five Scottish-English pairs, across the six conversations (in three tasks), there is on average ~7 mins of talk-time per speaker per conversation, based on the acoustically-aligned orthographic transcriptions. The spot-the-difference task provides the most speech, with an average of nearly 2 mins talk-time. The map task conversations are just over 1 minute, and the story recall task average is shorter, at around 40 secs. In the map task, the information-giver typically talks more than the follower (around twice as much), with more variety of structure and content. In the spot-the-difference task, however, things are more equalised. For some pairs there are phases within the conversation in which one speaker takes the lead, followed by an exchange of role; and there can be an asymmetry, in which one or other speaker is more loquacious overall. The story-retelling task tends to have longer contiguous speaker phrases, in both roles.

The average duration of the spontaneous shadowing is on average ~2 mins, and is more evenly split between the roles. The follower speaks faster, for less time, and often attempts to speak during the gaps between the longer phrases of the spontaneous speaker.

The read passage, Comma, provides on average ~2½ mins of speech. The average total for these varied connected speech tasks is therefore just over 14 mins of talk-time.

In the four most naturalistic collaborative discourses (map task and spot the difference), speakers overlap and leave silences during their interaction in a very natural-seeming way. Mutual silences vary, but total, on average, up to 1 min, while the amount of overlapping speech varies too, but on average is ~10secs per discourse.

In all the connected speech tasks, the number of distinct word types is, on a first analysis, ~1,800. The token count is over 20,000. The 10 most common word types (token counts between ~1,400 down to ~300) are: *the*, *and*, *a*, *to*, *of*, *I*, *it*, *in*, *that*, *yeah*. At the other extreme, the great majority of lexemes occur in only six tokens or fewer.

Qualitatively, the discourse sections sound naturalistic. Casual speech lexis like *uh-huh*, *like*, *so*, *ok*, are frequent, as are laughter and prosodic and pragmatic indications of relaxed and natural interaction. In addition to dysfluencies and speech errors, there are over 600 filled pauses, about evenly split between *uh* / *eh* and *uhm* /*ehm*.

The speech rate in the read passage, on the basis of the 520 scripted words, is ~4 words/sec. Speech rates in the unscripted tasks is far more variable.

5. Show-n-Tell

In the session at Interspeech we will exemplify some of the key phenomena that the DoubleTalk Corpus can address. One of the main aims of the corpus is to provide natural and robust data on variability in articulation and articulation-acoustics mappings, to augment existing datasets of read sentences and words. The discourses will also allow researchers to examine natural prosodic phrasing, and how pragmatic, informational and conversational factors interact with articulation. The shadowing data shows characteristics of a fast speech style, and has syntactic, lexical and phonetic errors. We will therefore exemplify its naturalness in a range of tasks.

It is, moreover, possible to approach listener behavior as a topic in its own right. For example, the relative stillness of the articulators during active listening (i.e. intra-discourse speaker silence) can be examined. More importantly, since the EMA data of the participants is synchronized, real conversational timing can be explored from an articulatory perspective. In conversation, audible and silent cues both exist to turn-taking activity. Some turn-taking activity, for example, appears to occur without any audible reflex at all, when the participant starts to articulate, but then truncates the production without speaking. We will exemplify silent examples plus episodes of audible overlapping speech (back-channel, competition, turn-taking) that show synchronized overlap in preparation.

Some silent behaviors are conversational: dysfluencies, errors of gestural intrusion and prosodically conditioned articulations. Others are allophonic / segmental, like a tongue tip gesture during an apparently glottal-stop variant of /t/. Drawing on such examples, our demo will use AAA software & other audio-visual supports, including slow-motion movies, to exemplify silent and other covert gestures in their natural conversational habitat.

6. Acknowledgements

The facility was funded by EPSRC EP/E01609X/1 and EP/E016539. We would like to thank the speakers, and also our collaborators on the facility and corpus, a full list of whom form the complete author list in [1].

7. References

[1] Geng, C., Turk, A, Scobbie, J.M., Macmartin, C. et al. "Recording speech articulation in dialogue: Evaluating a synchronized double electromagnetic articulography setup", J. Phonetics. In press, 2013.

[2] Tiede, M., Bundgaard-Nielsen, R., Kroos, C., Gibert, G., Attina, V., Kasisopa, B., Vatikiotis-Bateson, E. and Best, C. "Speech articulator movements recorded from facing talkers using two electromagnetic articulometer systems simultaneously", Proc. of Meetings on Acoustics, Vol. 11, 060007. 2012.

[3] Articulate Instruments Ltd. "Articulate Assistant Advanced". Computer program, available from http://www.articulateinstruments.com/ 2013.

[4] Boersma, P. and Weenink, D. "Praat: doing phonetics by computer". Computer program, available from http://www.praat.org/ 2013.

[5] Baron-Cohen, A. and Wheelwright, S. "The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences", J. of Autism and Developmental Disorders, 34(2), 163-175, 2004.

[6] McCullough, J., Somerville, B. and Honorof, D.N. "Comma Gets a Cure", 2000.

[7] Wells, J.C. Accents of English I: An Introduction, CUP, 1982.

[8] Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. & Weinert, R. "The HCRC Map Task Corpus", Language & Speech, 34:351-366, 1991.

[9] Van Engen, K. J., Baese-Berk, M., Baker, R.E., Choi, A., Kim, M. and Bradlow, A.R. "The Wildcat Corpus of Native- and Foreign-Accented English: Communicative efficiency across conversational dyads with varying language alignment profiles." Language & Speech, 53(4): 510-540, 2010.