2011-01-27

Using Hotspots as a Novel Method for Accessing Key Events in a Large Multi-Modal Corpus

Catherine Oertel
Trinity College Dublin

Celine De Looze
Trinity College Dublin

Brian Vaughan
Technological University Dublin, brian.vaughan@tudublin.ie

See next page for additional authors

Recommended Citation

Oertel, C. et al. (2011) Using Hotspots as a Novel Method for Accessing Key Events in a Large Multi-Modal Corpus. *New Tools and Methods for Very-Large-Scale Phonetics Research Conference,* University of Pennsylvania, United States of American, January 28-31.

Using hotspots as a novel method for accessing key events in a large multi-modal corpus

Catharine Oertel[1]*, Céline De Looze*[1]*, Brian Vaughan*[1]*, Emer Gilmartin*[1]*, Petra Wagner*[2]*, Nick Campbell*[1]

[1]Speech Communication Laboratory, Trinity College Dublin, Ireland
[2]Fakultät für Linguistik und Literaturwissenschaften, Universität Bielefeld, Germany
oertelgc@tcd.ie

Abstract

In 2009 we created the D64 corpus, a multi-modal corpus which consists of roughly eight hours of natural, non-directed spontaneous interaction in an informal setting. Five participants feature in the recordings and their conversations were captured by microphones (room, body mounted and head mounted), video cameras and a motion capture system. The large amount of video, audio and motion capture material made it necessary to structure and make available the corpus in such a way that it is easy to browse and query for various types of data that we term primary, secondary and tertiary. While users are able to make simple and highly structured searches, we discuss the use of conversational hotspots as a method of searching the data for key events in the corpus; thus enabling a user to obtain a broad overview of the data. In this paper we present an approach to structuring and presenting a multi-modal corpus based on our experience with the D64 corpus that is accessible over the web, incorporates an interactive front-end and is open to all interested researchers and students.

Index Terms: large-scale multi-modal corpus, corpus design, phonetics, interactive platform

1. Introduction

In recent years, the rapid development of speech technology has facilitated the use of extremely large corpora of speech in linguistic and socio-linguistic research. While read and strongly controlled conversational speech may still retain their importance for phonetic research, interest is now also turned to more spontaneous and less controlled recordings. Arguably, the fewer constraints placed on conversational speech, the more naturalistic the data set is. Examples of such naturalistic corpora are Switchboard [1], Call-Home [2] and the Childes corpus [3]. As human interaction is multi-modal in nature, studying it entails the consideration of its visual and auditory modalities. Examples of such multi-modal corpora are the AMI corpus [4] and the Bielefeld Speech and Gesture Alignment Corpus (SaGA) [5]. In the latter one, for instance, participants were asked to give route directions in order to study the interplay between speech and gestures. The D64 corpus also aimed to capture natural interaction for study in relation to different modalities, but without restricting the interaction to specific tasks such as map tasks, general role plays in meetings or formal work group meetings. The purpose was to collect speech which was not domain and/or function specific. Consideration was given to the level of control exerted over the recordings and its effect on the type of interaction between participants.

2. Our objectives

The objective of this paper is to discuss the logical and coherent structuring and presentation of a large multi-modal corpus with a web based interface. This is described in the context of the D64 corpus, created in Dublin in November 2009 [6]. First we give an overview of the D64 corpus, briefly detailing the methods used to create it. Then we discuss methods to organise and structure large amounts of multi-modal data. Finally, we examine the methods of creating a web-based interactive platform.

3. Overview of the D64 corpus

The D64 corpus was recorded over two successive days in a rented apartment, resulting in a total of eight hours of recorded data [6]. The apartment was chosen for its relaxed atmosphere in contrast to a laboratory setting. Comfortable chairs and sofas were placed around a coffee table. The apartment offered enough space for participants to move around and they were even given the possibility to eat and drink whenever they wanted. In the last session participants consumed alcohol. Five participants took part on the first day and four on the second. Three of the participants were male and two female. They were colleagues and/or friends (with the exception of one naive participant), ranging in age from early twenties to early sixties. The conversation was not directed and ranged widely over topics both trivial and technical. On both days, video, audio and motion capture data were recorded using video cameras, microphones and a motion capture system.

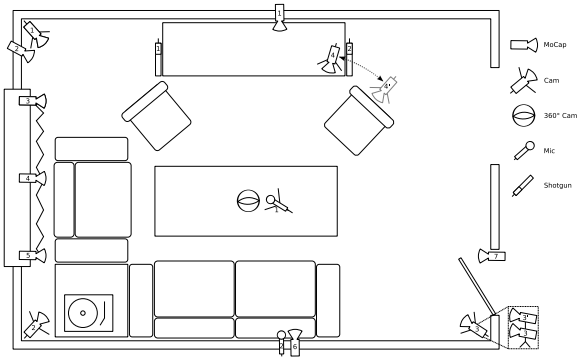

Figure 1: Room plot of the recording setup for the D64 corpus.

Figure 1 details the recording setup for the D64 corpus. As can be seen in Figure 1, the cameras were strategically placed

to cover all seats in the room. The 360 ° camera was placed in the middle of the seats to allow for all participants to be captured. The vinyl record turntable in the left hand corner was used to synchronise the motion capture cameras placed around the room. In addition to the microphones indicated in figure 2, participants wore head and/or body mounted microphones. A total of eight audio files and ten video files are available for Session 1 (see figure 2). Manual post-recording synchronisation was necessary as not all media devices were started at the same time.

FINAL_NAME	event time	local event time	ASSET_START	ASSET_STOP	Session
S1SHOT2		01:54:18	10:48:11	13:05:15	1
S1SHOT1		01:54:18	10:48:11	13:05:15	1
S1P5		01:54:18	10:48:11	13:05:15	1
S1P3P2		01:54:18	10:48:11	13:05:15	1
S1P2P3		01:54:18	10:48:11	13:05:15	1
S1P1		01:54:18	10:48:11	13:05:15	1
S1C4SOUNDRIGHT		01:54:18	10:48:11	13:03:15	1
S1C4SOUNDLEFT		01:54:18	10:48:11	13:03:15	1
S1C4002		02:33:27	10:09:02	13:03:15	1
S1C4001		01:00:25	11:42:04	13:03:15	1
S1C3004		off	12:45:53	13:12:44	1
S1C3003	00:15:00	00:15:00	12:27:29	12:45:53	1
S1C3002		off	12:09:06	12:27:29	1
S1C3001		off	11:50:42	12:09:06	1
S1C2002		00:02:57	12:39:32	13:01:06	1
S1C2001		off	11:50:18	12:39:32	1
S1C1002		off	12:02:33	12:23:20	1
S1C1001		off	11:34:34	12:02:33	1

Figure 2: Media Files Session 1.

4. Organising the D64 corpus

We structured the D64 corpus in three related layers: primary, secondary and tertiary data (see figure 3). We consider primary data to be the video, audio and motion capture files. The secondary data consists of the meta data (i.e. information relating to the participants, the interaction, the recording and the equipment). Tertiary data corresponds to multi-modal annotations of the corpus data: prosodic, gesture, discourse annotation etc.

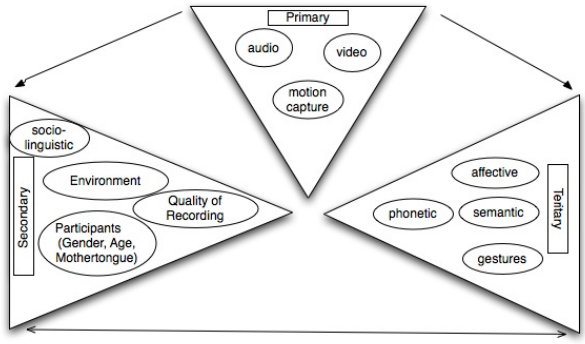

Figure 3: Primary, Secondary and Tertiary data.

4.1. Naming convention and synchronisation of audio and video files

In the case of the D64 corpus, both naming convention and synchronisation between the different files was conducted in 4 separate stages (see Figure 4). This was necessary because of the numerous different audio and video sources used. In a first step, criteria for the naming of the media files were chosen, with the final naming schema being as follows: the session during which the file was recorded, the format of the data (audio or video), and which person was recorded in the file. In a second step, all audio and video files were set into time relation to a master file. In a third step, all files were time coded in real time and converted into a common file format.

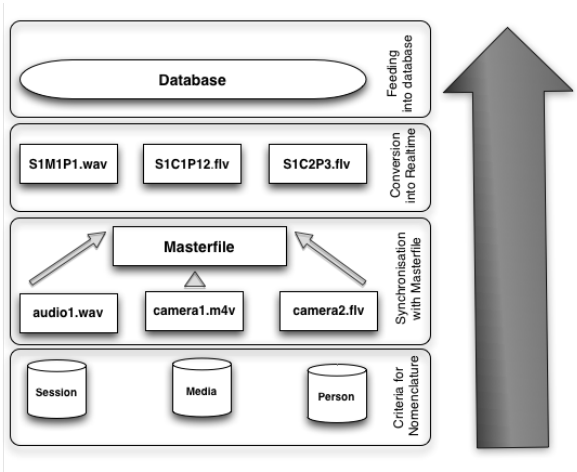

Figure 4: Outline of the corpus structure.

4.2. Metadata

Documentation of the metadata is an important aspect of corpus development. Types of metadata include, details of the level of interaction constraint (task-based, scripted dialogue, role play etc.), the recording equipment, the studio set up, and the participant information. Annotation of this type allows researchers to make informed decisions regarding the analysis of the data. In the context of the D64 corpus, social interaction took place between three high-ranking male academics and two female students. Within this interaction group, two of the male participants' native language was English while German, Dutch and Swedish were the native languages of the other participants. The whole conversation was held in English. Depending on the purpose of the study this type of information may be of interest. Various initiatives exist for annotating metadata: the Dublin Core Metadata Initiative (DCMI) [8] ; the Open Language Archive Community (OLAC) [9] ; the ISLE Metadata Initiative (IMDI) [10] and Mpeg7 [11]. Adaption of any particular metadata schema is dependent on the level of desired annotation and the type of the data obtained, with some schemas being more suited to particular types of data than others. Existing schemas are being tested in relation to the D64 corpus to ascertain the most suitable schema.

4.3. Annotations

The multi-modal nature of the D64 corpus means that various forms of annotations can be carried out e.g. phonetic, prosodic, affective, gestural and semantic. Annotation of corpus files can be carried out using a variety of open sources tools such as Praat [12], Wavesurver [13], ELAN [14] and Anvil [15]. Consideration must be given to relating this information to the primary (audio, video, motion capture files) and secondary (metadata) data. The following section discusses the data structuring.

5. Structuring the D64 data

Looking through a corpus in real time can be tedious and time consuming. In order to address this problem, a database can be used to structure the data in a meaningful and coherent way,

thus enabling relational querying and browsing to be carried out. Open source methods such as PHP [16] , HTML5 [17] and MySQL [18] allow the creation and use of a database in a cost effective, transparent and replicable manner. Structuring the data in such a way also offers the possibility of using a web-based front-end, built on top of the database, to provide a method for searching and querying the data: this facilitates wider access and use of the corpus data.

Within the corpus, hotspots are identified to enable a researcher to have a quick overview of key points in the conversational data. Following [19], we define hotspots as regions in which participants are highly involved in the discussion. A hotspot might be a location where people disagree, laugh simultaneously or where the dynamics of the conversation change noticeably.

The annotation schema designed for hotspots in our work assumes that listeners are generally able to detect hotspots without necessarily knowing what the current topic of the conversation is. Therefore, our annotation schema ought to rely on as little semantic pragmatic interpretation as possible. Compared to many multi-modal annotation schemas, our annotation schema only differentiates between two concepts - social distance and arousal [6]. The annotation schema proposed was assessed and presented at the Pink Cost 2102 conference in Budapest in 2010. See [7] for more information.

In addition to using hotspots, the data can also be queried in a simple or highly-structured manner. A simple query may consist of searching for the phonetic annotation of audio files, for a particular participant or all data related to all male participants in a corpus. Likewise a user might be interested in a more general search such as all data related to all male participants in the corpus. A highly-structured query might consist of searching for phonetic data relating to two different participants in a particular file (video, audio or motion capture) at a particular time.

The ability to query the data in this manner requires that the various forms of annotation be uploaded and parsed to the database. This can be achieved through an upload form on the browser front end, with the caveat that the structure and type of file for the various different forms of annotation must be standardised.

Collecting large amounts of audio and video and motion capture data requires a large amount of storage space. The D64 corpus is approximately 120 gigabytes in size. This is a problem when presenting and making accessible the data via a web-based interface. While the World Wide Web is a rapidly evolving medium, high resolution video and audio files must be compressed for reliable playback. The Mp3 format was developed for use on the web and is the defacto web audio standard: while lossy in nature [20], and not ideal for phonetic analysis, it is ideal for listening to audio files on the web. Likewise, the Mp4 video codec [21] is the de-facto web video standard and is supported across all major web browsers. The H.264 Mp4 codec in particular provides a good compromise between picture quality and level of compression [21]. Higher quality audio and video files are made available for download for offline playback and annotation.

6. The D64 web-based interactive platform

Sharing linguistic resources is essential to progress in interdisciplinary research. For example the CRDO-Aix (a Resource Center for the Description of Oral; [22]) provides academics with the possibility to share resources online. In a similar manner to the CRDO, we want to make the D64 corpus available on the web. We argue that a corpus that was originally devised for studying natural conversation, can be used both as a research tool and serve as a teaching resource. A browser based front end enables users to access the data and utilise it for their own ends. Furthermore, programing platforms such as Adobe Flex [23], Microsoft Silverlight [24] and the open source HTML 5 protocol [17] allow for the creation of Rich Internet Applications that can be used as interactive corpus interfaces and annotation tools [25].

In order to preserve work in progress but also because of ethical considerations, different levels of access have to be set up. Similarly to the CRDO, a tiered access system, which enables control through the use of a registration and password system, is put into place. Users can be given access to only some of the recordings and annotation data, while full access is restricted to a small number of researchers.

For a given user group of students, it is possible to use part of the D64 corpus to test and develop their hypotheses through the novel user-friendly database design available on the internet. The web based nature of the corpus and the database backend allows a level of interactivity not usually available with large corpora. Work is being carried out on methods to enable students to contribute their annotations to the D64 corpus, while being able to examine those annotations made by other users.

Web-based interfaces and annotation tools can be used to crowd-source corpus annotations [26, 27] and can provide a simpler corpus interface method [28]. Web-based annotation tools provide significant advantages over platform dependent, offline annotation tools, allowing large numbers of annotators to work in parallel and collectively on the same consistent corpus [29, 30].

While tools like ELAN [14] and Anvil [15] have been made available to display several video sources, work is being undertaken to provide a greater level of functionality and degree of interactivity for the students to annotate and visualize the D64 corpus data and annotation. Figure 5 displays our web interface design.

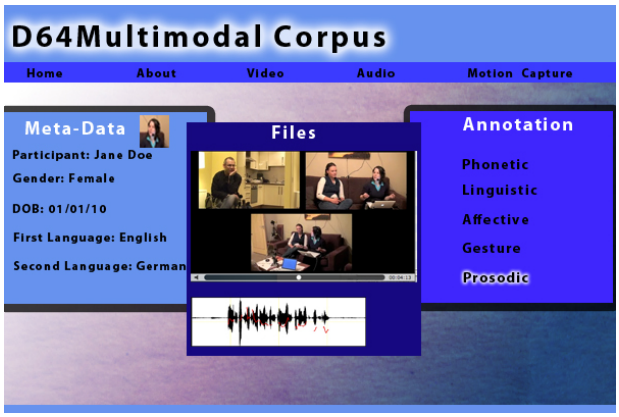

Figure 5: Web interface design.

7. Conclusion

This paper has examined methods of organising and structuring a multi-modal corpus based on our experience with the D64 corpus. We have discussed how we structured this corpus in three related layers, for primary, secondary and tertiary data.

Moreover, we have addressed issues on naming conventions and sychronisation of audio and video files, documentation of the metadata and other forms of annotation based on the corpus such as phonetic, prosodic, affective, gestural or semantic annotations. We have shown how a database can be used to structure the data in a meaningful and coherent way. This structuring enables a web based front end to be developed, thus providing wider access for students and researchers to the corpus data. This novel approach to corpus creation and distribution offers the possibility to query the data both in a simple and highly-structured manner. We also discussed the use of hotspots to give users a broad overview of key events in the D64 corpus. Furthermore, it was argued that a web based interface provides a convenient method of corpus annotation whilst providing wider access to the recorded data, thus facilitating the collection of large scale (crowd sourcing) annotations. The development of speech synthesis and related technologies, along with the development of more intuitive computer interfaces is often reliant on large readily available data sets.

8. Future Work

Work is continuing on the development of the back-end database and an easy to use web-based interface. Feasible methods of parsing offline annotation data are being examined, as are the various annotation file types and structures.

9. Acknowledgements

This work has been supported by grants to Nick Campbell from the Visiting Professorships & Fellowships Benefaction Fund from Trinity College Dublin, and the Kaken B Fund for Advanced Research from the Japanese Ministry of Information, Science &Technology and also Science Foundation Ireland (SFI), Stokes Professorship Award 07/SK/I1218. Catharine Oertel is supported by the Irish Research Council for Science, Engineering &Technology - EMBARK scholorship. This work was undertaken as part of the FASTNET project - Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) 09/IN.1/I2631. Finally, we thank Fred Cummins, Jens Edlund and Nike Stam for their collaboration.

10. References

[1] Godfrey, J., Holliman, E.C., McDaniel, J., "SWITCHBOARD: telephone speech corpus for research and development", IEEE International Conference on Acoustics, Speech and Signal Processing, Vol 1.: 517–520, 1992.

[2] Kingsbury, P., Strassel, S., McLemore, C. and McIntyre, R., "CALLHOME American English Transcripts", Linguistic Data Consortium, Philadelphia, 1997.

[3] MacWhinney, B., "The CHILDES Project: Tools for Analyzing Talk. Volume 1: Transcription format and programs. Volume 2: The Database.", Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[4] Carletta, J., "Unleashing the killer corpus: experiences in creating the multieverything AMI Meeting Corpus", Language Resources and Evaluation, vol. 41, n. 2: 181–190, 2007.

[5] Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H., "The Bielefeld Speech and Gesture Alignment Corpus (SaGA)", In M. Kipp, J.-C. Martin, P. Paggio and D. Heylen (Eds.), LREC 2010 Workshop: Multi-modal CorporaAdvances in Capturing, Coding and Analyzing Multimodality, 2010.

[6] Oertel, C. and Cummins, F. and Campbell, N. and Edlund, J. and Wagner, P., "D64: A Corpus of Richly Recorded Conversational Interaction", Proc. LREC 2010. Workshop: Multimodal Corpora–Advances in Capturing, Coding and Analyzing Multimodality. 2010.

[7] Oertel, C., Wagner, P., and Campbell, N., "Identification of cues for the automatic detection of hotspots.", Proceedings of the The pink cost 2102 international conference on analysis of verbal and nonverbal communication and enactment: the processing issues Budapest: http://berber.tmit.bme.hu/cost2102/: 67–68, 2010.

[8] Core, Dublin, "Dublin Core Metadata Initiative", http://dublincore.org/specifications/, 2009.

[9] OLAC, "Open Language Archives Community home page", http://www.language-archives.org/, 2008.

[10] Eagle/ISLE, "IMDI Metadata Domain", http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI240467#, 2007.

[11] MPEG "Mpeg-7 Overview", http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm, 2004.

[12] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", 2006.

[13] KTH Royal Institute Of Technology, "Wavesurfer Homepage", http://www.speech.kth.se/wavesurfer/index2.html, 2010.

[14] Hellwig, B., "Elan- linguistic annotator", Retrieved from http://www.lat-mpi.eu/tools/elan/, 2009.

[15] Kipp, M.,"Anvil- a generic annotation tool for multi-modal dialogue", Proceedings of the 7th european conference on speech communication and technology (Eurospeech), Aalborg, Denmark:1367–1370, 2001.

[16] Group, The P H P, "Home page of The PHP Group", http://php.net/index.php, 2009.

[17] W3C "HTML 5 Specification", http://www.w3.org/TR/html5/, 2010.

[18] Sun-Microsystems, "My SQL Home Page", http://www.mysql.com/, 2009.

[19] Wrede, B. and Shriberg, E., "Spotting hot spots in meetings: Human judgments and prosodic cues", Eighth European Conference on Speech Communication and Technology, Geneva, Switzerland:2805–2808, 2003.

[20] Hacker, S., "MP3: The Definitive Guide", O'Reilly (1st ed.) 400p., 2000.

[21] Koenen, R., "MPEG-4 Overview", Forum American Bar Association: 1–79, 2002.

[22] Bel, B. and Blache, P., "Le Centre de Ressources pour la Description de l'Oral (CRDO)", 2006.

[23] Adobe, Adobe Flex Home Page, http://www.adobe.com/products/flex/, 2009.

[24] Microsoft, Microsoft SilverLight, 2010.

[25] Cullen Vaughan, B., Mc Auley, J., and Mc Carthy, E., "CorpVis: An Online Emotional Speech Corpora Visualisation Interface", Springer Berlin / Heidelberg, vol 5887, 169–172: 2009.

[26] Green, N., Breimyer, P., Kumar, V. and Samatova, N., "WebBANC: Building Semantically-Rich Annotated Corpora from Web User Annotations of Minority Language", Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA, Jokinen, K. and Bick, E. (eds), (NEALT) Northen European Association for Language Technology: 48–56, 2009.

[27] Tarasov, A., Cullen, C. and Delany, S.J., "Using crowdsourcing for labelling emotional speech assets", w3.org, n09, http://www.w3.org/2010/10/emotionml/papers/tarasov.pdf: 1–5, 2010.

[28] Apostolova, E., Neilan, S., An, G., Tomuro, N. and Lytinen, S., "Djangology: A Light-weight Web-based Tool for Distributed Collaborative Text Annotation", Proceedings of the International Conference on Language Resources and Evaluation, 2010.

[29] Draxler, C., WebTranscribe: An Extensible Web-Based Speech Annotation Framework, Lecture Notes in Computer Science: Text, Speech and Dialogue, vol 3658: 61–68, 2005

[30] Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J. and Ave, N., "Collecting Image Annotations Using Amazon Mechanical Turk", Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Datawith Amazons Mechanical Turk, June, 139–147, 2010.