Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 46 (2017) 175–178

www.elsevier.com/locate/csl

Towards the next generation of speech tools and corpora

Christoph Draxler*, Jonathan Harrington, Florian Schiel

Institute for Phonetics and Speech Processing, Ludwig Maximilian University of Munich, Munich, Germany

This special edition picks up the theme 16 years after Bird and Harrington (2001) of current developments in software tools for processing speech and language data. The main objective now is much as it was then: to design and make freely available tools that are independent of the research task and computing environment for creating, annotating, querying, and analysing data from extensive speech and language corpora that often originate from a wide range of disciplinary perspectives in the speech sciences. Since then, the advances in web-technology combined with greater processing power and data storage have resulted in what Schiel and Kisler (2017) refer to as a paradigm shift in developing software tools for speech and language corpora (see also predictions already made in this direction by Draxler (1997) through the development of his WWWTranscribe tool).

There have also been major changes in the nature of the data that are analysed in at least two ways. Firstly, in terms of the disciplines that the development of speech and language corpora brings together: as Cassidy and Estival (2017) note, these encompass at the very least ’acoustic phonetics, speech technology, natural language processing, lexicography, socio-linguistics and linguistics more generally, but also include aspects of psychology and musicology’. Secondly, in terms of quantity: recording, annotating, analysing and processing the 3.1 million infant vocalisations referred to in Beckman et al. (2017) of nearly 1500 day-long recordings (at 2 h of processing time per day) would have been scarcely possible back in 2001.

The major paradigm shift is that emerging web technologies and standards foster the development of web services that complement or even replace specialised stand-alone tools for processing speech and language data. The advantages of a web-based approach as also outlined in Cassidy and Estival (2017), Schiel and Kisler (2017), and Winkelmann et al. (2017) are clear: because the tools are run within a web-browser, they are platform-independent, thereby removing both the need for programmers to update tools separately on each platform and for users to reinstall the tools locally after system upgrades. There are also some disadvantages: ubiquitous and high-speed network access is necessary, access to the local machine is restricted for security reasons, there still are browser incompatibilities, and there are legal issues when sensitive data have to be transferred via the network (see a brief discussion in Schiel and Kisler 2017).

Web-based solutions have nevertheless led to an increasing democratisation of speech and language analysis: on the one hand, the availability of high-quality open source software for visualisation and analysis has made it easier for researchers and students of speech and language processing to process large amounts of speech and language data; on the other, the familiarity and ubiquity of the web has led to a standardised work flow and data management that facilitates access to and reuse of speech data. Advances in hardware technology, reduced cost, and the adaptation of novel techniques also contributed to this democratisation.

The contributions in this volume reflect the technological and methodological developments since 2001. Although each article has a specific thematic focus, the individual articles are linked by common challenges, technological requirements and data structures.

* Corresponding author.

E-mail address: draxler@phonetik.uni-muenchen.de (C. Draxler), jmh@phonetik.uni-muenchen.de (J. Harrington), schiel@phonetik.uni-muenchen.de (F. Schiel).

http://dx.doi.org/10.1016/j.csl.2017.05.007
0885-2308/ 2017 Published by Elsevier Ltd.

Pouplier and Hoole (2017) analyse the dynamics of vowel-to-vowel coarticulation using ultrasound technology. According to them, using ultrasound has the potential to lead to a democratisation of speech production research which 'has so far largely been restricted to comparatively few technically high-powered phonetics labs in the world'. This development has been accompanied by greater availability in the last 10−15 years of large corpora of speech physiology data and tools for processing them. In 2001, it is very unlikely that the last two editions of *Computer Speech & Language* summarised in Cassidy and Estival (2017) would have contained studies that make use of audio recordings combined with MRI scans and a method for synthesising speech from articulography data.

Readily accessible techniques such as ultrasound, the types of technological advances in processing image data as presented in Pouplier and Hoole (2017), and the greater availability of corpora containing speech physiology data are critical for advancing many of the issues that are discussed in Beckman et al. (2017): in particular how children manage to achieve phonological stability during language acquisition, taking into account that the growth and development of the articulatory system during acquisition has such a marked influence on the acoustic signal and its perceptual interpretation.

Developing a generalised data model for representing, accessing, and analysing speech and language data is a central theme in many of the papers in this edition. As Winkelmann et al. (2017) note, a generalised model is designed to overcome the problem of speech and language corpora that have been created with highly specialised tools and that lack a common interface. The idea behind a generalised data model is that there is a set of procedures that define the set of possible speech and language databases i.e. principles to which all such databases, their signals and annotations conform. Working out what such principles are is of course a major challenge—it includes for example having a model that can represent any kind of annotation tagged to any of the possible multi-media speech signals that could occur for any speakers in any speaking style and in any language. The real challenge is therefore to strike a balance between generality and domain specificity: at one extreme, the mathematical graph formalism covers all known types of signals and annotations and has well-understood formal properties (Bird and Liberman, 2001), but in practice these properties are of limited value. At the other extreme, the exclusively time-based segments in many annotation tools prevent modelling the complex non-linear relationships in representing speech and language. The data model proposed by Winkelmann et al. (2017) combines the practical requirements of querying and analysing speech data with well-defined data structures.

One of the many reasons why it is important to strive towards a generalised data model (and here there is a clear continuity with the ambitions from several of the papers in the Bird and Harrington (2001) special edition on a similar theme) is reproducibility. This important theme is taken up in Cassidy and Estival (2017). If researchers continue to build and analyse corpora with their own highly specialised techniques, then the scientific transparency and hence validity are compromised because of the resulting difficulties of replicating exactly the same experiment in another laboratory. Their concern in developing the Alveo22 Virtual Laboratory, a web based repository for language data, is to address this issue by designing a system for sharing both data and reproducible workflows. The advantage of doing so, they suggest, is that researchers using Alveo can discover tools from each other's disciplines—enabling for example techniques from speaker diarisation technology to be made available to historians engaged in social history research.

Both Fromont (2017) and Winkelmann et al. (2017) take up the long-standing challenge of how to develop a generalised annotation model. Both systems—LaBB-CAT in Fromont (2017) and Emu in Winkelmann et al. (2017)—allow annotations to be represented independently of their time stamps. This is argued in both papers to be important because annotations need not be anchored in time: a timeless annotation is appropriate, for example, for representing the canonically present but often physically absent schwa in the production of the syllable-final nasal in the second syllable of *sudden*. As Fromont (2017) argues, a further advantage of freeing annotations from time stamps is to give expression to the idea that the serial order of annotations is not necessarily predictable from their temporal order (an idea that is also central to theories of autosegmental phonology—Goldsmith, 1976). Both annotation models in Fromont (2017) and Winkelmann et al. (2017) also take up the challenge of how to represent intersecting hierarchies of annotations so that annotations at a lower tier can be parsed in multiple ways at a superordinate tier. LaBB-CAT provides some primitives for querying annotations. The query-language in Emu remains very similar to that in earlier versions (Cassidy and Harrington, 2001; Harrington, 1993) but with a further extension to include regular expressions. Both LaBB-CAT and Emu are open-source and browser-based. The emphasis in LaBB-CAT is on providing a common integrated representation for annotations that might have been created with different tools and in a diversity of annotation formats. A major new development in Emu is its seamless integration with the R programming environment so that in contrast to earlier systems, the creation, querying, and analysis of speech corpora are all operations that can be carried out from R from which a web-based application for viewing and further annotating speech data can be launched. Web-based solutions have the further advantage of enabling multiple researchers possibly located at geographically different sites to update the annotations of a commonly shared database.

The issue of speeding up the very labour-intensive task of annotation is addressed from different perspectives by various studies in the special edition. Schiel and Kisler (2017) present a web-based extension to the Munich automatic segmentation system MAUS that has been developed over a number of years (starting with Schiel, 1999) for the segmentation of speech based on an acoustic and a phonological representation of the language. In this extended system, the output following an upload of paired orthographic and signal data via a web-browser is a segmentation of phonetic annotations in Praat, Emu, or BAS-Partitur formats. Since MAUS forms part of a tool-chain with both the SpeechRecorder (Draxler and Jänsch, 2004) system for capturing speech data and Emu (Winkelmann et al., 2017), users can obtain parameterised signals and annotations for analysis in R often within minutes following speaker recordings. Further public web-services described in Schiel and Kisler (2017) are text-to-phoneme conversion, syllabification, optimal text alignment and speech synthesis.

The types of processing discussed in Schiel and Kisler (2017) cannot be readily extended to most types of prosodic annotations for various reasons: firstly, because there is nothing equivalent in sentence-level prosody to a known orthographic transcription within which prosodic annotations can be forced to align; secondly, because there are not enough sizeable databases to build reliable, statistically-based training models; and thirdly, because the consistency in annotating phrase boundaries, prominence, and utterance-level tones even across expert transcribers is much lower than for segmental annotations and typically often around 70−80% (e.g. Grice et al., 1996). Cole et al. (2017) suggest a possible way out of this impasse by using crowd-sourcing techniques to mark only very broad prosodic annotations that have some meaning also for lay-persons: specifically, whether or not a transcriber hears a word as prominent and a boundary between two words. The advantage of this crowd-sourcing approach is that it rapidly yields vast amounts of prosodically annotated data. Cole et al. (2017) show that annotations derived from this method are at least as reliable as those from student annotators; and perhaps more importantly that lay-persons and experts in prosodic mark-up draw upon the same sets of acoustic and contextual cues in annotating the data prosodically. Beckman et al. (2017) are confronted with a similar type of problem of obtaining rapidly large quantities of reliably annotated child speech data. Consistently with Cole et al. (2017), they also propose using a roughly-based annotation methodology in order to provide a framework within which annotation refinements could be subsequently made. They suggest deploying ’rough diarisation methods’ that can now quite reliably pick out infants’ utterances from a conversational background; and then subsequently bringing to bear the small number of databases in which children have been recorded longitudinally for annotating child data more narrowly.

Both Cole et al. (2017) and Schiel and Kisler (2017) take up the important issue of extending automatic annotation to multiple languages. The segmentation system in Schiel and Kisler (2017) which currently works on 19 languages as diverse as German and Hungarian is planned for extension in the next 2 years to all European languages and the five most frequently spoken languages in the world. Cole et al. (2017) point out that the type of rapid prosodic annotation that they have demonstrated may well be the most efficient way of obtaining much needed prosodically annotated data from multiple and above all less well studied languages. One of the reasons why the method of Cole et al. (2017) is so important in obtaining prosodic databases of less well studied languages is because of the great difficulty—if not impossibility—of training a sufficient number of experts in their accurate prosodic annotation, especially since accurate intonational transcription necessitates near native-like proficiency in the language. Finding ways to augment rapidly child speech databases with annotations that have linguistically and phonetically useful associations with speech signals will be essential to continue with the various new insights into language acquisition in early childhood that analysing multilingual speech data provides (Edwards et al., 2015).

The papers in this special issue all show not only the great advances that have been achieved in the handling of digital empiric speech data since 2001 but also demonstrate that modern speech processing techniques and high-bandwidth access to the Internet will allow even the non-technical trained scientist to apply state-of-the-art analysis and modelling in the near future.

Looking backwards, we can state that the five major topics addressed in the special edition of Bird and Harrington (2001), namely *scalability*, *adaptability*, *multi-layered structures*, *querying* and a *discourse-prosody interface* have been tackled successfully, but important challenges remain. Looking forward, we can discern two new major topics: the long-term *availability* of tools and resources, and *sharing* or *collaborative use* of data. Speech data repositories are growing, funding agencies now often require a data management plan (leading to more data being shared), and—maybe most importantly—researchers can now exchange data or access the same data easily or even informally. These new topics raise critical questions of sustainability, intellectual property rights and the legal protection of researchers who distribute speech data as well as their informants which need to be answered in the near future.

Acknowledgements

The editors are grateful to the 20 colleagues who undertook timely and insightful reviews of the submissions.

References

Beckman, M., Plummer, A., Munson, B., Reidy, P., Edwards, J., 2017. Methods for eliciting, annotating, and analyzing databases for child speech development. Computer Speech and Language. 45, 278–299.

Bird, S., Harrington, J., 2001. Speech annotation and corpus tools. Speech Commun. 33, 1–4.

Bird, S., Liberman, M., 2001. A formal framework for linguistic annotation. Speech Commun. 33, 23–60.

Cassidy, S., Harrington, J., 2001. Multi-level annotation in the EMU speech database management system. Speech Commun. 33, 61–77.

Cassidy, S., Estival, D., 2017. Supporting accessibility and reproducability in language research in the Alveo Virtual Laboratory. Computer Speech and Language. 45, 375–391.

Cole, J., Mahrt, T., Roy, Josef, 2017. Crowdsourcing for prosodic annotation. Computer Speech and Language. 45, 300–325.

Draxler, C., 1997. WWWTranscribe - a modular transcription system based on the world wide web. In: Proceedings of the Eurospeech, Rhodes, pp. 1691–1694.

Draxler, Chr., Jänsch, K., 2004. SpeechRecorder − a universal platform independent multi-channel audio recording software. In: Proceedings of the Language Resources and Evaluation Conference, Lisbonpp. 559–562.

Edwards, J., Beckman, M., Munson, B., 2015. Cross-language differences in acquisition. In: Redford, M. (Ed.), The Handbook of Speech Production. John Wiley & Sons, New Jersey.

Fromont, R., 2017. Toward a Format-Neutral Annotation Store. Computer Speech and Language. 45, 348–374.

Goldsmith, J., 1976. Autosegmental Phonology. Massachusetts Institute of Technology, Dept. of Foreign Literatures and Linguistics. Ph.D. Thesis http://hdl.handle.net/1721.1/16388.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J., Batliner, A., 1996. Consistency in transcription and labelling of German intonation with GToBI. In: Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia, pp. 1716–1719.

Harrington, J., 1993. The MU+ system for corpus based speech research. Computer Speech and Language 7, 305–331.

Pouplier, M., Hoole, Ph., 2017. Öhman returns: New horizons in the collection and analysis of articulatory data. Computer Speech and Language.

Schiel, F., 1999. Automatic phonetic transcription of non-prompted speech. In: Proceedings of the International Congress of Phonetic Sciences, San Francisco, pp. 607–610.

Schiel, F., Kisler, Th., 2017. Multilingual processing of speech via web-services. Computer Speech and Language. 45, 326–347.