Resources for Speech Research: Present and Future Infrastructure Needs

Lou Boves[1], *Rolf Carlson*[2], *Erhard Hinrichs*[3], *David House*[2], *Steven Krauwer*[4], *Lothar Lemnitzer*[3], *Martti Vainio*[5], *Peter Wittenburg*[6]

[1] Department of Language and Speech, University of Nijmegen
[2] Speech, Music and Hearing, KTH, Stockholm
[3] Seminar für Sprachwissenschaft, Universität Tübingen
[4] Utrecht institute of Linguistics UiL OTS, Utrecht University
[5] Department of General Linguistics/Speech Sciences, University of Helsinki
[6] Max Planck Institute for Psycholinguistics, Nijmegen

L.Boves@let.ru.nl, rolf@speech.kth.se, eh@sfs.uni-tuebingen.de, davidh@speech.kth.se, steven.krauwer@let.uu.nl, lothar@sfs.uni-tuebingen.de, martti.vainio@helsinki.fi, Peter.Wittenburg@mpi.nl

Abstract

This paper introduces the EU-FP7 project CLARIN, a joint effort of over 150 institutions in Europe, aimed at the creation of a sustainable language resources and technology infrastructure for the humanities and social sciences research community. The paper briefly introduces the vision behind the project and how it relates to speech research with a focus on the contributions that CLARIN can and will make to research in spoken language processing.

Index Terms: speech research, resources, standardization

1. Introduction

1.1. The CLARIN Project

CLARIN is a joint effort of over 150 institutions in Europe, aimed at the creation of a sustainable language resources and technology infrastructure for the humanities and social sciences research community. Its objective is to create a federation of existing digital archives in Europe that contain language based data (e.g. text, speech, multimodal) and tools (e.g. for acquisition, annotation, exploration) and that offer researchers unified single sign-on access to this material, as well as to web based language and speech analysis services to perform both simple and complex operations on the data, using proven tools and technologies.

At this moment CLARIN has received funding from the EC (4.1 M€) and from 23 participating countries (13 M€) for a preparatory phase (2008-1010), during which the foundations for the infrastructure will be laid. This preparatory phase in the form of an FP-7 project with a consortium of currently 33 partners in 23 countries) will comprise technical work (the full technical specifications of the service infrastructure and the construction of an experimental prototype to validate the specifications), requirements collection (from the prospective users), language and speech technology work (surveys of what exists in terms of resources and technologies, agreement on standards for exchange and interoperability), legal issues (access and authentication, IPR and licensing, business models), training and dissemination, and, last but not least governance and funding (reaching an agreement between funding agencies in the participating countries on the future organisation and governance of the infrastructure and sustainable funding). Work along all these lines is ongoing, and is expected to be finished at the end of 2010. We will then move on to the next phase, the construction phase, which will last 3-5 years. Finally, by 2015 a well-funded and sustainable infrastructure will be in place. .

CLARIN's target audience is the humanities and social sciences research community at large and from the CLARIN perspective all languages are equally important, irrespective of size or economic importance.

In this paper we will focus on aspects related to speech and multimodal resources and related technologies.

1.2. The CLARIN Vision

The vision of CLARIN is to turn existing, fragmented language resources and tools into accessible and stable services that any user -- not just technology experts -- can access and use for research and development purposes. Such user-friendly infrastructure would include services that cut across various language modalities: spoken language, written language, and mixed modalities that combine text, speech, gestures, and video. For example, a historian may want to analyze interviews with Jewish immigrants from different geographic regions in Europe about their reasons to move to Israel and combine these with the data from the Shoah Foundation Institute at the University of Southern California[1]. Once the recordings, which are stored in different repositories, have been retrieved, the historian might need to obtain an automatic transcription of the speech. While such a project may only be feasible in the future [2], it is necessary to think already now about the tools, services and facilities that would be needed: a meta-data scheme that describes the data and allows sophisticated matching of queries, documents and processing tools, in addition to on-line documentation detailing any legal and/or ethical restrictions on the use and distribution of the individual data records.

If the basic functionalities and services exist at all today, their use is often mired by IPR and copyright restrictions as well as by proprietary data formats that are often not compatible with one another. Moreover, current tools and resources are have been created by engineers for technology specialists, rather than with humanities scholars in mind, who want to be able to use tools and services without having to first go through extensive upfront training.

The goal of CLARIN is to unite this fragmented landscape of language resources and tools and to construct a persistent and stable infrastructure that researchers can rely on. Our vision is

[1] http://college.usc.edu/vhi/

that the resources for processing language, the data to be processed, the processing tools, as well as appropriate guidance, advice and training will be made available on-line in a distributed network from the user's desktop. CLARIN will make this vision a reality: user will have access to a growing set of resources, tools and guidance from distributed knowledge centers. Most importantly, access will be via a single sign-on. Thus, researchers will be relieved from the need to obtain access to multiple, independent resources in future eScience projects what will combine data from hitherto unrelated resources. Last but not least, CLARIN aims to make available the supercomputer resources that are needed for future speech and language research in a service oriented architecture based on secure grid technologies.

1.3. Speech research is special

There are at least two important questions regarding speech research that are related to the creation of a future research infrastructure available to the researchers: 1) are there important lines of research that have been hindered by the lack of resources and tools and, 2) what are the gaps in the infrastructure responsible for the hindrance? It is well known that the answer to 1 is "yes"; there are indeed important lines of research that cannot be conducted by the community at large. This is, of course, true with regard to every branch of science, but the situation regarding speech research is perhaps more dire than in some other (in some ways comparable) fields such as biology. However, the situation can only be fixed once the real needs of the community have been identified. Which leads to the second question regarding the lines of research that the new infrastructure should enable. Rather than trying to fix holes that do not exist, the work for enabling future research should start from real needs. Although identifying current needs is easier than predicting future needs, the task is in no way trivial. Rather than surveying a set of experts from various fields as to "what they would like to have", we must establish where the science of speech is today and how it relates to neighboring branches of science.

Speech as a phenomenon is studied in many branches of science separately. Therefore, a vision regarding the common infrastructure should not be imposed by one powerful sub-community. Nevertheless, a coherent vision is required to make sense of the plethora of needs indicated by the diverse communities. In this paper we introduce one possible vision regarding the current and future needs for a common European infrastructure for speech research.

CLARIN intends to include all existing resources in one virtual repository. Nevertheless, some thought should be spent on how to prioritize the work. Some data and tools require more work than others and there is a trade-off between quantity and quality and differences regarding the ease of acquisition.

2. Current situation

The speech community is large and very diverse and the needs for infrastructure are not even obvious for everybody involved. If we consider speech technology in the sense of training ASR or synthesis there is a rather good platform to distribute the resources through the ELDA or LDC for the large languages. The Basic Language Resource Kit (BLARK) [10] description of the current situation and wishes for the future is a rather general view and very much inspired by the need of resources for speech technology. However, if we look at more humanities types of issues such as child-directed speech, accent shifts, dialogs in different environments and second language learners, we have a very different and fragmented situation. Currently it is very much the researchers' own task to collect the data or try to do research on a corpus collected for some other purpose. To be able to share data oriented to the humanities would be a big step for many researchers. To appreciate the richness in humanities oriented speech research one can look at the International Congress of Phonetic Sciences (ICPhS), which was last held in Saarbrücken in 2007.

In addition to the more traditional representation of phonetics research topics such as prosody, acoustics, production and perception, two relatively new trends can be seen in the research areas presented at the ICPhS in 2007. These trends can be characterized as studies involving phonetic variation and phonetic detail, and studies involving expressive speech and non-verbal aspects of spoken language. The area of phonetic detail and variation was represented in the keynote talk about phonetic detail and interaction [13]. There was also a special session on modeling fine phonetic detail. In addition, attention to phonetic variation and fine phonetic detail appeared as a theme in many contributions in a number of research areas such as dialect and social variation, L2 acquisition, and cross-linguistic variation. The use of data-bases and speech corpora to attempt to capture variation was prevalent in many of these contributions. The research area of expressive speech and non-verbal aspects of speech including visual speech has attracted a rising interest among speech researchers in recent years. This area was represented by the keynote talk about acoustic and visual aspects of verbal and non-verbal communication [1]. Problems of obtaining multimodal data involving expressive and emotional speech are not trivial and the need for collaboration and sharing of data is obvious. Finally, it is of interest to note that in the proceedings of the Interspeech Conference in Antwerp in 2007, there were about 5500 instances of the word “database” or “corpus” and 3000 the word “prosody” or “prosodic” while in the papers of ICPhS 2007, the ratio was reversed. Although these ratios are different, it is clear that databases are becoming increasingly important not only for speech technology but also for basic research in phonetics.

2.1. Tools & Resources

2.1.1. Speech Tools

Despite substantial progress over the last decades we are still far from having speech tools that are so powerful and robust that they can be used safely and effectively in a way similar to office productivity tools. Often, research projects cannot suffice with standard usage of the most popular features of the tools. This problem is certainly aggravated by the fact that speech tools tend to be designed by engineers for a specific purpose, which is not conducive to good user interface design and documentation. It would be interesting to understand why ESPS/XWaves+ was not taken up by the speech community, while Wavesurfer [13] and PRAAT [3] are extremely popular.

For everything related to speech synthesis (from building multimodal teaching support systems to generating stimuli for perception experiments) there is the Festival toolkit [4]. Yet, to use Festival for intonation research one needs to dig below the surface, and doing so requires knowledge about speech synthesis as a technology. For many things related to speech recognition there is HTK [5]. Since HTK is being used by thousands of researchers world wide, there is extensive experience with standard applications. Yet, everybody who is new to the tool (and to automatic speech recognition) will experience a substantial learning curve, the more so if this person is not working close to colleagues who can share their own experience.

Tools such as Festival and HTK are not panacea; there remain tasks for which these tools are not the best option. This was the reason why the Dutch-Flemish STEVIN Program (arguably the best example of a BLARK initiative for a tier 2 language [6]) has funded a project aimed at bringing the HMM-based automatic speech recognition software developed at ESAT in Leuven into the public domain [7]. This project, called SPRAAK, involved 55 person months, mostly senior researchers. The experience gained in the SPRAAK project emphasizes the need for sustainable maintenance and support, as well as for training of researchers who can not rely on extensive knowledge about automatic speech recognition. To name just one example: providing a tool that can perform automatic alignment and segmentation of spontaneous speech corpora and that is easy to use for phoneticians is still an open issue.

Thus, CLARIN is facing a major challenge when it comes to the ambition to make speech tools available to the eHumanities community at large. The ambition to create an infrastructure reminiscent of the university libraries and the computer and network infrastructure that we are used to cannot be fulfilled at no cost. Without the support of such an infrastructure the CLARIN ambition is tantamount to hoping that it will always be possible to find researchers who are able to get the lasts bits of functionality and performance from existing tools and who have the time and the resources to collaborate in projects of other researchers at essentially no costs. What is crucial, nevertheless, is that CLARIN will be populated by materials and tools according to the users' needs from the beginning. This, on the other hand, depends on local funding in each part-taking country.

2.1.2. Corpora

Invariably surveys of existing speech corpora yield lists of recordings, often accompanied by orthographic transcriptions, of a wide range of speech styles in a wide range of languages. Yet, equally invariably, many recordings are only available on tape, with transcriptions as hard copies. Even if transcriptions are available in computer readable format and if great care was spent in maximizing their quality, turning such a corpus into a resource that can be used with automated tools and can be shared with researchers in other laboratories may be surprisingly expensive. As an example, take the corpus of ten 90 minute Dutch spontaneous conversations, collected by Ernestus just before the turn of the century. As is usual with conversational speech the chunks into which conversations are split up comprise multiple turns. This implies the presence of overlapping speech in most chunks. For automatic processing, e.g. word level alignment or phonetic transcription, cross-talk due to overlapping speech is a serious problem. Cross-talk was also a major issue in the spontaneous conversations in the Spoken Dutch Corpus (CGN) [9], but in the orthographic transcriptions of the CGN this (and similar) issues were anticipated, resulting in the decision to use chunks no longer than 3 seconds. While removing chunks with cross-talk from the former corpus would amount to scratching more than half of the speech, in the CGN we could rescue the bulk of the speech. For automatic alignment canonical phonemic representations of all words in the corpus are needed. It appeared that of the 9500 word types more than 1000 were missing in the Dutch TST-Lexicon that contains some 350,000 words [8]. While we expected some OOV words (because Dutch allows the creation of compounds and because part of the conversations were role playing involving brand names created for the purpose), the proportion found seemed exceedingly high. It appeared that part of the OOVs were due to spelling issues (for example the use of hyphens in compounds). In CGN spelling issues were avoided by imposing a transcription protocol that limited transcribers to using words in a (growing) list of forms canonized by a lexicographer.

The effort for making the corpus collected by Ernestus amenable to automatic processing involved a combination of expert labor and the deployment of tools that are fit for re-use (possibly as web services). CLARIN will help to make those tools accessible and usable for researchers who cannot rely on the assistance of local speech technology expertise.

2.2. Examples of emerging research

In speech research, there is a general trend towards integration of different disciplines around specific problems centered on speech. These groups tend to offer new and fresh points of view to old questions. These projects range from modeling speech recognition and perception in new ways to modeling speech production with new mathematics, as is done in a new project on articulatory speech synthesis at the Institute of Mathematics of the Helsinki University of Technology. There, speech production serves the purpose of mathematical research in addition to the intended results in the form of new articulatory models [11]

Emotional and affective speech and their synthesis as well as recognition is a field that has gained ground during the last decades. What projects that address new problems have in common is their requirements regarding data; the data they use (or propose to use) is very difficult to obtain. They range from controlled interactional data including non-speech sounds to affective speech and three-dimensional articulatory data. On the other hand, the projects are in the position to provide the community with entirely new tools and ways of handling data. Thus, it is essential that all kinds of research groups take part in building the common research infrastructures.

From a more technological point view, speech-based services are becoming part of our everyday life and we expect them to be efficient and robust and that it will be easy to understand their function. Still, the dialog systems have far from human-like behavior due to many known and unknown factors. Recently, exciting progress has been made pushing the research on human-human and human-machine interaction forward. Efficient modeling of error-handling [14] is an example of such progress taking into account how different factors, such as dialog context, acoustic probability and cost of misunderstanding, influence the treatment. Data-driven approaches [15] to building dialog models are other examples of new methods to model discourse. The steps towards conversational systems make modeling of for example turn-taking [16], dysfluencies, and emotion [17] important research fields. Progress in these areas is very much dependent on the accessibility to well structured and annotated dialog corpora including multimodal and multiparty interaction [1]. The expense of developing such corpora is very high since for example each utterance can not be studied independent of its dialog context and thus the size of the corpus needs to be large to achieve representative coverage. As we take the next steps, leaving the well-structured dialog path, towards modeling of human conversation, the CLARIN infrastructure will be a necessary type of tool to share large corpora of human interaction.

3. Conclusion

In this paper we have very briefly presented the CLARIN project, which aims at creating a language resources and technology infrastructure for the humanities and social sciences. This infrastructure should provide access to text, speech and multimodal resources and tools to support and innovate humanities and social sciences research for all languages. Here we have focused on aspects related to speech and multimodal resources and technology, and we have made a number of observations.

We believe that on the basis of existing or emerging speech technology it should be possible to provide humanities researchers with powerful and innovative research instruments that allow them to combine data from different sources. The role of CLARIN is to create a context in which existing expertise and resources can be brought together and made available to the community. This requires a significant effort towards the adoption of interoperability and representation standards, as well as adaptation of the existing technologies for a non-technical audience of users. This cannot be achieved without financial support from the funding agencies in parallel with support for the creation of the infrastructure itself. Experience from recent projects shows that the effort required for adaptation of existing tools and resources to common standards or to new applications is difficult to overestimate.

The coverage in terms of existing resources is quite uneven: some languages cannot even meet the emerging BLARK standard for minimal coverage, whereas at the same time new, cutting edge approaches to speech and multimodal technologies require highly specialized resources in large quantities. Here too a special funding effort is needed, to go hand in hand with the development of CLARIN: completion and maintenance of the BLARK in parallel with the development of new types of resources to support emerging technologies. This paper highlights examples, such as articulatory speech synthesis, handling emotion, human-human and human-machine interaction to illustrate the problem and the lack of resources available. The production of such resources is expensive, and only a well-coordinated effort in the production of these resources and efficient instruments to share and maintain them can ensure their timely availability to the research community. CLARIN can play an important role in the coordination of these activities, but for the implementation coordinated support from the funding agencies is necessary.

A third observation, especially applicable to speech resources is that where until recently speech resources were mostly looked upon as massive bulk data collections used for training of speech tools and applications, there is now a shift towards more content-based processing of speech data. This will no doubt lead to new avenues in speech research from which the humanities and social sciences scholars can reap enormous benefits. Where in the past historical language material used to be mostly textual the 20th century has brought the research community enormous quantities of audio and video material that are waiting to be explored, provided it is made accessible and comes with appropriate exploration tools.

The main message of this paper is that CLARIN can be an extremely useful facility for the research community, but only if the construction of the infrastructure is accompanied by national and international programmes aimed at populating it with the resources and tools that the researchers will need.

4. References

[1] Granström, B., & House, D. (2007). Inside out - Acoustic and visual aspects of verbal and non-verbal communication. Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, 11-18.

[2] Inkpen, D., Alzghool, M., Jones, G.J.F. and Oard., D.W. (2006) Investigating cross-language speech retrieval for a spontaneous conversational speech collection. In Conference on Human Language Technologies and the North American Chapter of the Ass'n for Computational Linguistics, New York

[3] Paul Boersma & David Weenink (2005) Praat: doing phonetics by computer. http://www.praat.org/

[4] Clark, R.A.J., Richmond, R. King, S. (2004) Festival 2 - build your own general purpose unit selection speech synthesiser. In Proc. 5th ISCA workshop on speech synthesis.

[5] Young, S., Evermann, G., Kershaw, D. Moore, G., Odell, J., Ollason, D., Povey, D. Valtchev, V., Woodland, P. (2002) The HTK Book (For HTK Version 3.2). Cambridge University Engineering Department.

[6] D'Halleweyn,E., Odijk, J.,Teunissen, L., Cucchiaini, C. (2006) The Dutch-Flemish HLT Programme STEVI: Essential Speech and Language Technology Resources, Proc. ELREC-2006, pp. 761-766.

[7] Demuynck, K., Duchateau, J., Van Compernolle, D.,Wambacq, P. (2000) An Efficient Search Space Representation for Large Vocabulary Continuous Speech Recognition. Speech Communication, volume 30, pp. 37—53.

[8] Schuppler, B., Ernestus, M., Scharenborg, O, Boves, L. (2008) Preparing a Corpus of Dutch Spontaneous Dialogues for Automatic Phonetic Analysis, Proc. Interspeech-2008, pp. 1638 – 1641.

[9] Oostdijk, N., Broeder, D. (2003) The Spoken Dutch Corpus and Its Exploitation Environment. In Proc. 4th Int. Workshop on Linguistically Interpreted Corpora (LINC-03). Budapest, Hungary.

[10] Krauwer, S. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap (2003) Proceedings of 2nd International Conference on Speech and Computer (SPECOM2003)

[11] Hannukainen, A. and Lukkari, T. and Malinen, J. and Palo, P. (2007) Vowel formants from the wave equation. J. Acoust. Soc. Am. 122 (1)

[12] Local, J. (2007). Phonetic detail and the organization of talk-in-interaction. Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, 1-10.

[13] Sjölander, K. and Beskow, J. (2006) Wavesurfer software http://www.speech.kth.se/wavesurfer/

[14] Skantze, G. (2007). Error Handling in Spoken Dialogue Systems - Managing Uncertainty, Grounding and Miscommunication. Doctoral dissertation, KTH, Department of Speech, Music and Hearing.

[15] Lemon, O. and Pietquin, O. (2007) Machine Learning for Spoken Dialogue Systems, Proc. Interspeech 2007.

[16] Edlund, J., Heldner, M., & Gustafson, J. (2006). Two faces of spoken dialogue systems. In Interspeech 2006 - ICSLP Satellite Workshop Dialogue on Dialogues: Pittsburgh PA, USA.

[17] Shriberg, E. E. (2005). Spontaneous Speech: How People Really Talk, and Why Engineers Should Care. Proc. Eurospeech, Lisbon. Portugal.