ON PRINCIPLES OF PHONETIC ARCHIVING: FROM PALEO-PHONETICS TO MODERN SPEECH DATA MANAGEMENT

Christoph Draxler[1], *Jürgen Trouvain*[2]

[1]*Ludwig-Maximilians University, Munich, Germany,*
[2]*Saarland University, Saarbrücken, Germany*
draxler@phonetik.uni-muenchen.de

Abstract: Repositories and archives play a dual role in phonetics and speech research. Looking forward, they provide current and future accessibility to resources being compiled now. Looking backward, they now provide access to resources compiled in the past. We argue that by adhering to a few general principles early in the process, speech resources can be compiled with little extra effort in such a way that they can easily be deposited in repositories for future accessibility. We further argue that because of the particular nature of phonetics and speech research data, archives and archival units specifically dedicated to this type of data are needed for the long-term preservation of and accessibility to these resources. In this paper we outline these general principles and describe the actors and their responsibilities. Although in theory these principles should be clear, in practice it is a) necessary to remind researchers and focus their attention to the issues of long-term preservation of phonetic research data, and b) for this task we advocate for dedicated institutions linked to the research labs.

Keywords: Speech database, repository, data curation, metadata, archive

1 Introduction

In empirical research data is a key factor. In phonetics, speech science, and technology development, compiling data is time-consuming and thus expensive. It requires trained personnel, manual processing and transcription, and very often high-tech equipment. It is thus clear that such data collection effort should be honoured by storing the data in a repository or archive so that it can be shared and made available for future research, or to replicate results. However, we observe that speech data collected in research projects are not always deposited in archives or made available to others. In our experience, many research projects in our fields suffer from the following main problems with respect to data sharing and long-term availability:

1. Speakers have not provided consent for data storage and sharing, or the legal status of existing data is unclear.
2. The metadata describing a speech data collection is insufficient.
3. Curating data for and depositing data in repositories is often neither considered to be scientific work nor is sufficient funding available.
4. Long-term availability and accessibility do not map well to short-term project funding.

Finally, there is another, albeit less-confessed reason for not depositing data in an archive: researchers fear that by putting data in a repository, they will give up a competitive advantage or have their data taken away from them.

In this paper, we first outline the specific data requirements of phonetics and speech research. We then describe the tasks, actors, and legal requirements involved in data collection, archiving and exploitation.

2 Data in Phonetics

A speech database as it is used in phonetics, speech science, or technology development, consists of three types of data [3]:

- *Primary* or *raw data* are the original recordings
- *Secondary* or *dependent data* are signal data derived from or annotations of primary data
- *Metadata* are descriptions of the speech database contents, speaker-related data, logs, legal texts, etc.

Primary data may not be modified or changed. Secondary data and metadata may be re-computed, edited, corrected, or adapted.

2.1 Primary data

In phonetics, speech science, and technology development, primary data consist of time-aligned audio, video, or sensor data. Video data cover not only common video, but also e.g. ultrasound, X-ray, fMRI, EEG or pupillometry and perception data (judgment, categorizations, reaction times etc.). Sensor data yield raw physical measurements such as x,y-position, acceleration and speed, physiological measurements, etc. Such data are produced by for instance X-ray microbeam, electropalatography, laryngography, electro-magnetic articulography, or eye-tracking.

2.2 Secondary data

Derived signal data, e.g. spectra, f0 or formant tracks or vocal tract area functions are computed from primary data. In general, the derived signal data depends on the particular algorithm used. Changes in algorithms may thus lead to different results even for the same primary data.

Annotation is the process of assigning a symbolic label to a continuous signal in a categorisation process. For speech, it is recommended to have at least three levels of annotation: an orthographic transcript of the recorded speech, a broad phonemic or canonical representation, and a narrow phonetic segmentation and labeling, see e.g. [5: 162]. Other types of annotation concern for example prosodic, syntactic, discourse, and semantic information. Technically, annotations are generally organised in tiers, with a tier containing only elements of a given type, and links between the elements within and across tiers, see e.g. [1, 8] for details.

2.3 Metadata

Metadata describes the structure and general content of a speech database. Ideally, this description is human-readable so that it can be made available in a web-based catalog, and machine-readable for general purpose or dedicated search engines. Examples of metadata standards are Dublin Core [7] or CMDI (Component Metadata Infrastructure, [2]). Furthermore, metadata also comprises examples of information sheets, consent forms, etc. used in the speech databases.

2.4 Analogue data

Currently, signal and annotation data is generally recorded, created, and stored in digital form. However, there exist historical recordings and annotations on analogue media, e.g. audio and video cassettes or manuscripts. These media need to be digitized for further processing and archiving. For an example, see the account of the digitization of the X-Ray Film Database in [6].

3 Archives and Repositories

Archives and repositories are often used as synonyms. In our view, a repository contains only digital research data, whereas an archive may also contain analogue data. Traditionally, an archive is a well-structured long-term store of historical records. The main purpose of an archive is to preserve artefacts – documents, works of art, technology – in their original state, to document them, and to provide access to these artefacts. Repositories basically serve the same purpose as archives, but they focus on the data proper.

Archives and repositories play a dual role in science. This role relates to time: looking forward, they provide accessibility to data that is being compiled today. Looking backward, they provide understanding of and accessibility to data compiled and technology used in the past. Thus, we consider phonetic archiving as a task ranging from "paleo-phonetics" to modern speech data management.

Ideally, archives and repositories are institutions with long-term funding, and a focus on a given domain, e.g. speech data and technology for research. The particular data types in phonetics and speech research require dedicated archives and repositories with a thorough understanding of the research field to assess the relevance of technology and the collected data.

3.1 Phonetic archives and repositories

Some well-known speech data archives are the phonogram archives in Zurich, Vienna, or Berlin, which exist since the early 20th century and contain valuable historic audio recordings on original media and the corresponding recording and playback devices, e. g. early dialectal or language documentation recordings on wax cylinders. Interestingly, these archives were originally planned not only to be used by researchers, but also for the general public. Besides these archives, the media and technology industry have built their own archives, e.g. of their broadcasts.

However, access to these archives is restricted, and their contents are in general not fully digitised. With the advent of digital speech processing and the increasing demand for contemporary speech data for technology development, repositories for speech data such as the Linguistic Data Consortium (LDC) in the US, the European Language Resources Association (ELRA) in Europe or the Bavarian Archive of Speech Signals (BAS) in Germany, to name a few, have been established. Originally planned to facilitate sharing and distribution of speech data for research and technology development, they now increasingly serve as archives – the TIMIT database, originally described in [4] and now maintained by the LDC, is already more than 30 years old. Admittedly, finding existing repositories and archives can still present a challenge.

3.2 Archive and repository users

Users of phonetic archives and repositories typically are interested in the historical aspects of speech communication, or they wish to re-use data that other researchers have compiled. Historical aspects range from understanding the techniques and the achievements of our forerunners to diachronic work, e.g. by comparing audio data from 100 years ago with those from today.

Re-using existing data – and possibly contributing to these data – allows researchers to perform their own analyses on existing data, to compare their results with those of others, and to augment the value of a speech database by adding new layers of annotation. Finally, many funding organisations now require that research data is made available after a project has ended.

Thus, researchers in the field face the following questions:

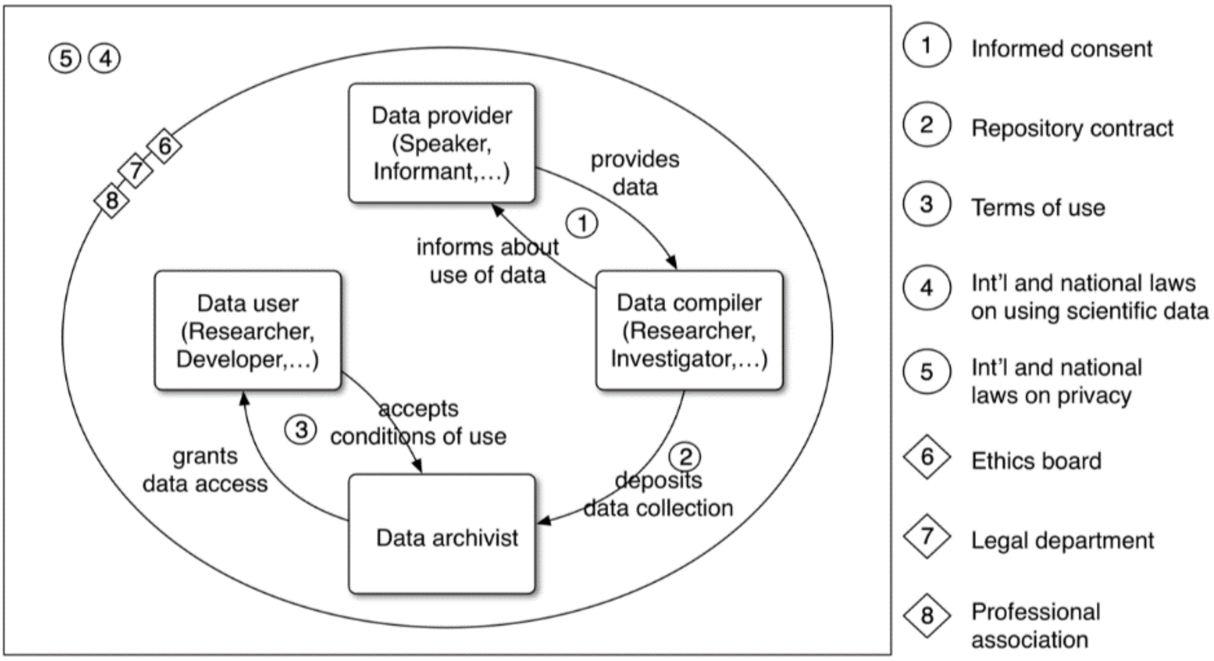
- Is there data "out there" that I can use for my work? This would really facilitate my work!

- I have to create a speech database. What should I care for in technical, legal and archiving terms?
- I've found old data in the lab which is possibly valuable to others. How do I proceed now?
- Collecting my data was laborious. How can I make them available beyond my project?
- The funding body (or a journal) explicitly asks for a data management plan. How can I do that in the best way?

These questions will be addressed in the following section.

4 Actors and Tasks

Figure 1 gives an overview of the actors and tasks relevant to compiling, depositing, and exploiting speech data.



5
4
6
7
8
Data provider (Speaker, Informant,…)
provides data
1
informs about use of data
Data user (Researcher, Developer,…)
Data compiler (Researcher, Investigator,…)
accepts conditions of use
3
grants data access
2
deposits data collection
Data archivist
1 Informed consent
2 Repository contract
3 Terms of use
4 Int'l and national laws on using scientific data
5 Int'l and national laws on privacy
6 Ethics board
7 Legal department
8 Professional association

Figure 1. Schema of the actors, documents, and the legal and professional regulations regarding the recording, archiving and usage of speech data.

4.1 Data compiler

The data compiler is a researcher who plans and performs a data collection. There is a clearly defined research question, which is formulated in a speech database specification in a project proposal.

It is recommended to consult the legal department (before starting the data collection) of the institutions involved and to request assistance in preparing the necessary documents such as consent forms, agreements on archiving data, etc. The legal department of course is bound by international and national laws, e.g. copyright and privacy regulations. It may be worth mentioning that many institutions (universities, scientific associations, clinical organisations) now have their own ethics committees, which also need to be consulted.

The data compiler is the legal owner of the data, and usually is the first to work with the data. In general, this person has the right to exclusive access to the data for a given period, e.g. the project duration. After this period, the data should be shared.

A special case are speech databases where the data compiler is unknown – such databases exist in many places. Often, these databases consist of valuable recordings, but there is no documentation of speaker consent on using or sharing the data. Again, the legal department should be consulted. It will base its judgment on the juridical risk involved with using, sharing or granting access to this data.

4.2 Data provider

The data provider – or speaker, informant, participant – is a person actually producing the speech data to be recorded. The data compiler informs the data provider about the purpose of the recordings, the intended use of the data, possible risks, and asks the provider for consent, usually in writing.

It is good practice to separate private information on a data provider, e.g. name, address, from socio-demographic data needed for the analyses, e.g. gender, age, hearing impairment, or regional background. Furthermore, the consent form should be short, written in easy to understand language, and it should allow answering the consent questions separately.

Typically, these questions are arranged from very specific to general consent: The data

- will only be used in the current project,
- will also be used outside the project in education,
- will also be shared for research, technology development, and education, with controlled access,
- will also be licensed for commercial purposes.

Note that the data provider also has the right to have her/his data removed from a database. This only concerns future use of the data. Clearly, the data provider has to trust the data compiler to apply standards of good practice and to comply with legal regulations.

4.3 Data archivist

The data archivist is responsible for the long-term storage of speech databases. This person receives data from the data provider, imports it into the repository, adds the metadata to the catalog and informs potential users about the new data. In general, the data provider remains the owner of the data. For the archive, the data must be in a format accepted by the archive or repository, and it must be accompanied by a contract governing the deposit of and access to the data.

The data archivist also has to make sure that the data stored survives technology changes. There are two solutions to this problem: a) continually migrate the data to new media and formats without data loss (e.g. through compression), and b) keep old hardware functional as long as possible; this may be facilitated by maintaining an updated list of sites that can process given materials. In the catalogs, updates of the data are reflected by versioning the data in the archive.

It is good practice to provide at least two modes of access to archives and repositories: browsing and searching. The first requires a human-readable web page, the latter either a human-readable search interface or a machine-readable application programming interface. Finally, the data archivist has to consult the data compilers on best practices of data collection and packaging.

4.4 Data user

The data user is a researcher, developer, or teacher who uses speech data for their purposes. This person searches for and retrieves data from archives and repositories, and must accept the conditions of use of a given archive or repository, e.g. to give credit to the data providers or to not re-distribute the data. It is good practice to report errors found in the data to the data archivist, and to contribute to a given speech database by extending the secondary data by new annotations.

4.5 Funding

The actors described above have different interests and time perspectives, and there is a mismatch between these interests and perspectives. Research is performed in short-term projects (typically, two to five years) by researchers with limited or permanent contracts. Research institutes provide an infrastructure for the projects; institutes often exist around 25 years (or shorter). Repositories and archives aim for long-term (longer than 25 years) data preservation.

It cannot be the task of researchers to provide long-term accessibility of their research data. Ideally, there exists a publicly funded institution linked or even within the research labs to serve as a repository or archive. As an alternative, part of the project funding is assigned to depositing data in an archive – thus, an archive continually receives funding for its services from short-term research projects. Many funding agencies currently require that projects define a data management plan to ensure data access for at least 10 years after a project has ended.

5 Conclusions

Archiving phonetic data is a topic that concerns virtually all labs working with phonetic data. Increasingly, journals and funding bodies require that research data is accessible for extended periods of time. We thus propose the following recommendations:

- research data should be shared,
- research data should be available for a minimum of 10 years after the end of a project,
- educate students on the importance of sharing data and make researchers aware of the merits of sharing data,
- make sure that there is a citable reference to the research data,
- when compiling research data, follow the guidelines recommended by recognised archives and repositories,
- set aside funding to prepare data for archiving and for the archiving itself,
- delegate the archiving task to someone with a permanent position, or to a dedicated archive or repository.

Apart from raising the awareness for these issues in our field, we hope that this paper is a helpful contribution to the development of phonetic archives in the phonetic and linguistic communities.

Acknowledgements

The authors would like to thank Bernd Möbius for very valuable comments on this paper.

References

[1] Bird, S., Liberman, M. 2001. A formal framework for linguistic annotation. Speech Communication 33(1,2), 23–60.

[2] Broeder, D., Windhouwer, M., van Uytvanck, D., Goosen, T., Trippel, T. 2012. CMDI: a component metadata infrastructure. Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR pp. 1–4.

[3] Draxler, C. 2008. Korpusbasierte Sprachverarbeitung - eine Einführung. Tübingen: Gunter Narr.

[4] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D. S., Dahlgren, N. 1986. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NIST.

[5] Gibbon, D., Moore, R., Winski, R. 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin: Mouton de Gruyter.

[6] Munhall, K., Vatikotis-Bateson, E., Tohkura, Y. 1995. X-ray film database for speech research. Journal of the Acoustical Society of America (98), 1222–1224.

[7] Weibel, S., Kunze, J., Lagoze, C., Wolf, M.1998. Dublin Core Metadata for Resource Discovery. IETF #2413. The Internet Society, September 1998.

[8] Winkelmann, R., Harrington, J., Jänsch, K. 2017. Emu-SDMS: Advanced Speech Database Management and Analysis in R. Computer Speech and Language.

URLS

1. www.phonogrammarchiv.uzh.ch/de.html
2. www.oeaw.ac.at/phonogrammarchiv/
3. de.wikipedia.org/wiki/Berliner_Phonogramm-Archiv
4. www.dublincore.org/specifications/dublin-core/dces/1998-09-01/
5. www.ldc.upenn.edu
6. www.elra.info/en/
7. www.bas.uni-muenchen.de/Bas/