

# Designing a Finnish Multimodal Speech Database System

Toomas Altosaar<sup>1</sup>, Mietta Lennes<sup>2</sup>, Manne Miettinen<sup>3</sup>, Mickel Grönroos<sup>3</sup>, and Matti Karjalainen<sup>1</sup>

<sup>1</sup>Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland

<sup>2</sup>Department of Phonetics, University of Helsinki, Finland

<sup>3</sup>CSC - Scientific Computing Ltd., Espoo, Finland

E-mail: Toomas.Altosaar@hut.fi, Mietta.Lennes@helsinki.fi, Manne.Miettinen@csc.fi,  
Mickel.Gronroos@csc.fi, Matti.Karjalainen@hut.fi

## ABSTRACT

Many different research groups in Finland use Finnish speech material as a basis for their studies. Since a general speech database is not available and there exist no common guidelines for collecting and annotating speech material, speech corpora end up being compiled in a variety of ways — often for just a single research purpose. This is both inefficient and inhibitory to interdisciplinary cooperation. To improve the situation, a project named "Integrated Resources for Speech Technology and Spoken Language Research in Finland" was initiated. The goal of the project is to design several exemplar and prototypical multimodal speech database systems that can serve as examples for these research groups, build a conforming but extendable infrastructure by selecting, building and refining necessary applications for speech annotation and database access, and to compile guidelines for annotation so that in the future, speech corpora can be shared readily.

## 1. INTRODUCTION

Several different research groups in Finland have collected a diverse variety of Finnish speech. These corpora come in the form of audio as well as multimodal recordings from one or more speakers covering speech from many different speaking styles. Small portions of these corpora have been annotated either manually or by semi-automatic means. For example, recordings may have related orthographic transliterations, phonetic labels and segment boundaries, linguistic descriptions, prosodic transcriptions, etc. Sometimes the annotations are stored independently from the recordings and in other cases they are linked. In some advanced systems the file system is hidden from the user altogether and interaction occurs between objects instead.

Several problems exist in the manner in which these recordings have been produced and annotated. First of all, common terminology is lacking between different groups thus hindering a free flow of interdisciplinary knowledge. Secondly, ad hoc solutions for annotation formats have frequently been applied making data exchange difficult. An additional incentive is related to the well-being of the language itself: Finnish is a small language and demands special attention with regard to speech technology

applications. If a system for common speech resources were available, multidisciplinary cooperation could be greatly enhanced thus improving our knowledge of spoken language.

To address these problems a project was launched in 2002 named "Integrated Resources for Speech Technology and Spoken Language Research in Finland". Funded by the Academy of Finland for three years, the project aims to find common guidelines and technical solutions for the collection, annotation, and distribution of digital Finnish multi-modal speech data. These goals can only be achieved by taking into account the requirements of all potential research groups dealing with speech. A majority of these groups have realized the need for a set of integrated resources and are directly involved in the project. Various fields of speech technology and spoken language research are represented: phonetics, linguistics, conversation analysis, engineering, etc.

During the three-year project, four exemplar databases will be collected and annotated according to the standards developed. Two of the databases will contain audiovisual speech and the common annotation scheme will be XML-compatible. User-friendly tools are being developed for both annotating and searching corpora either locally or across a network.

## 2. TERMINOLOGY AND RDF

One of the most difficult problems in designing a general and multimodal speech database system is that researchers coming from different fields do not have a common terminology for describing speech. International efforts have been taken to build standards for the collection and annotation of language corpora [1, 2] and for the representation of metadata, especially on the web, e.g., RDF [3]. The results of these projects provide valuable starting points for defining speech-related terminology.

### 2.1 TERMINOLOGY

In this project, the terminological foundations of what a speech database structure should be have been discussed. A workshop was held in 2002 to address the views of different interest groups. Our work has revealed that only a few

“standard” units that are frequently used in speech terminology have been agreed upon. For this reason, it has been decided that the number of fixed definitions for annotated units should be kept to a minimum. Also, existing international standards will be taken into account when building the basic conceptual framework for annotation.

Figure 1 shows some basic concepts: entities, metadata, and some annotations that might be included in a speech database. This type of conceptual map can be used to discuss and formalize the relationships between different types of data entities.

In practice, however, the level to which standardization of speech-related concepts and annotation practices can be carried out is limited. It must be recognized that different experiences and viewpoints will always produce different conceptual systems, which is also an important factor behind scientific progress. However, the diversity of concepts for describing speech leads to special technical requirements for the tools and data formats that are used in common speech databases.

## 2.2 RESOURCE DESCRIPTION FRAMEWORK

In order to promote — or even enforce — a (national) standard for speech annotation, RDF (Resource Description Framework; more specifically the RDF/XML syntax) was chosen as the format for annotation data files as well as for metadata about the signals being annotated. RDF enables centralized schemas on the World Wide Web (web) that define both (a) the **annotation units** (e.g., phones, syllables, words, or other annotated features, e.g., voice quality) that are formally supported, and (b) the required **metadata** of a media file (e.g., information about the speakers and the environment where the signal was recorded). An annotation editor and search system can then utilize these online schemas.

The Resource Description Framework (RDF) [3] was originally developed for the efficient processing and sharing of metadata on the web. It is intended to provide an abstract formal framework to describe things or resources without committing to any specific application domain or platform.

RDF allows many different sets of definitions to exist for the same types of resources. This can be seen as an asset since it allows the system to find structural similarities between RDF schemas that use totally different vocabularies. Thus, RDF makes it possible to exchange and share systems of definitions that can be partially modified or extended to suit different views and applications. These features also make RDF a feasible candidate for the document format of speech annotations. For instance, annotators will initially be forced to use a simple, common template or “core” schema to explicitly define the necessary concepts and annotations for a common database. Once this information is provided, any user may add or

override the definitions with concepts from any public or local RDF schema.

## 3. RDF-BASED ANNOTATION EDITOR

In order to promote the usage of common annotation guidelines developed in this project, researchers will require an annotation tool that is aware of these constraints. Annotation editors in general do not support reading or writing of RDF documents. Therefore, one of the goals of this project is to develop an RDF-based annotation tool for spoken language. This editor is able to read RDF schemas from URLs in remote locations and can enforce the annotation files that are created to follow such schemas.

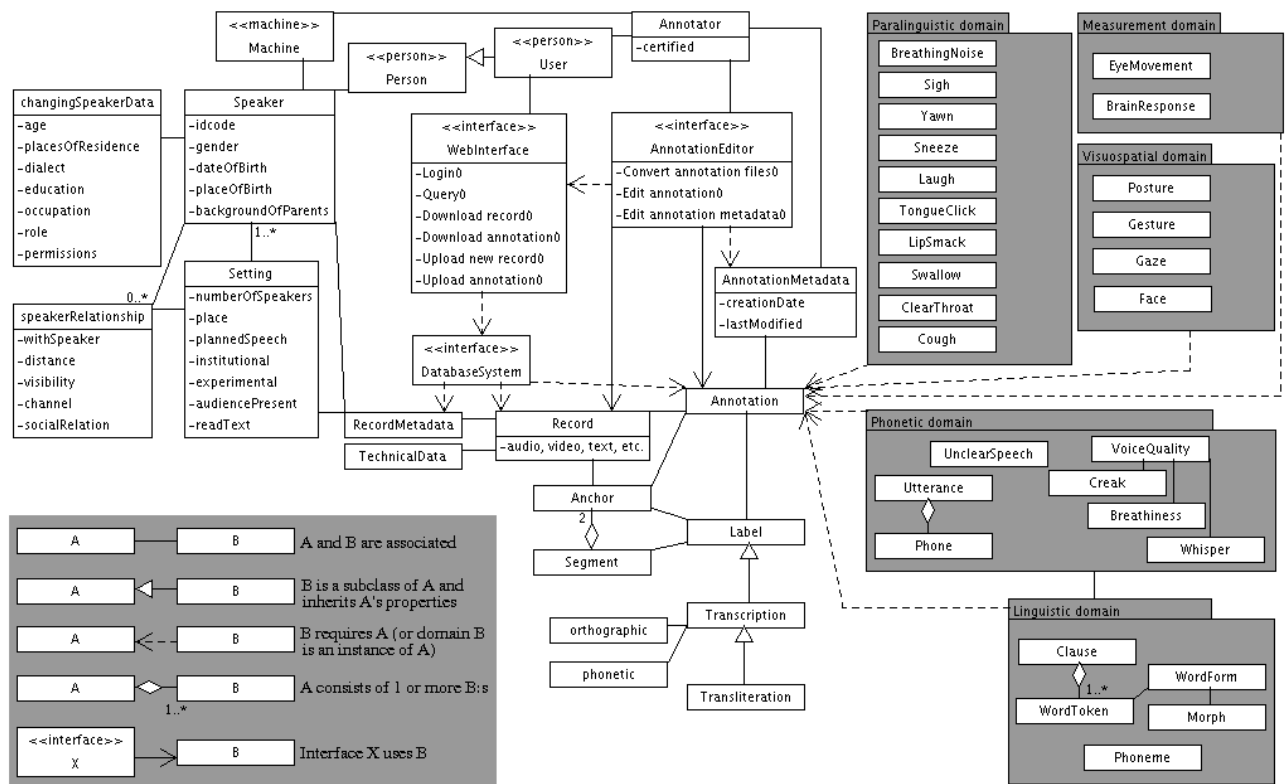
This means that official annotation schemas published in RDF can be located on a centralized server, e.g., the web server of the Language Bank of Finland. Schemas can be globally located while actual editor software runs on a users’ local machine. Therefore, actual instances of the annotation schemas, i.e., the files containing annotation, can be published on the web for use by other researchers, in an integrated and coherent manner. Furthermore, the user of the annotation editor never needs to deal with actual file names and similar technical details since an abstraction layer exists which hides low-level implementation issues.

The annotation editor is implemented in Python, a programming language that was chosen for a number of reasons. First of all, Python is to a large extent platform independent, which is a practical requirement when trying to connect all researchers in the fields of speech and speech technology. Secondly, Python comes with a fairly large and standard graphical user interface library. Finally, there is already a well-functioning open source library for sound visualization available, the *wsurf* library [4], written in Tcl but easily used from Python.

## 4. OBJECT MODELS OF SPEECH CORPORA

The object-oriented (OO) formalism is well suited to modeling annotated speech so that it can be searched and analyzed efficiently. For example, an OO perspective allows large speech databases in multiple languages to be queried using linguistically motivated structures [5].

First, terminology local to this paper is defined. Different types of speech related information, e.g., orthographic, linguistic, phonetic, prosodic, conversational, etc., can be seen to represent instances of some view or theory of language. A **domain** is defined as any one of these views or theories. In a phonetic domain, e.g., there may exist several different hierarchical **levels** of activity such as an utterance, word, or phone. A level in a domain contains actual instances of activity or events called **units**, e.g., a phone [a] or an F0 value. Finally, a **framework** associates any number of different domains and can be used to richly describe speech activity as well as explicitly map the



**Figure 1:** Possible relationships between different types of data entities in collections of speech.

relationships between different views or theories. A framework usually models one utterance or “chunk” of speech, e.g., a sentence. Frameworks can be linked to one another to model longer passages of speech activity.

RDF schemas resemble class hierarchies in the OO paradigm and therefore can be used to provide a natural starting point in describing speech related information. RDF can be used effectively to normalize the varied information in existing speech corpora as well, e.g., symbols used in different phonetic alphabets, e.g., IPA, Worldbet, SAMPA.

The goal is to compile speech recordings and annotations into efficiently searchable structures. These object frameworks are in practice created from the annotations by processing them with respect to the associated RDF schemas. For example, audio signals, phones, words, and phrases, are compiled into new **units** whose different features can be analyzed without loss of generality or efficiency. Units are first placed into **levels**, which in turn are assigned to a **domain** in some desired hierarchical order. Finally, a set of domains is grouped together to form a **framework**.

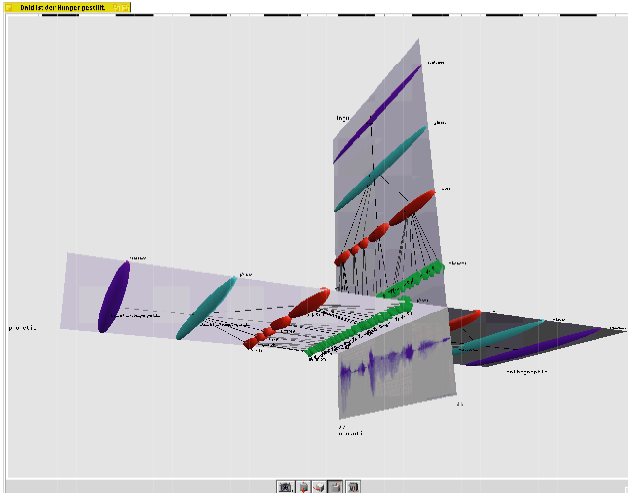
Relationships between units on similar or different levels and even in different domains are represented by **links**. This allows units to be aware of their local context, e.g., a

phone unit knows which units it is adjacent to, the syllable(s) or word(s) it is part of, or segments that may represent its substructure (e.g., closure and release for stop consonants). Figure 2 shows a framework that has been derived from a speech recording and its associated annotations. By exploiting object-class hierarchies and recursive queries, corpora can be accessed and studied in an intuitive and advanced manner.

## 5. DATABASE QUERY

Databases formed from speech corpora which have been modeled using frameworks can be efficiently queried by using a matching paradigm. First, a user defines a *template structure* which serves as a model for the query. For example, a typical query may be to find all [a] (IPA) occurrences in CVC syllables spoken by females in a set of multi-lingual or other non-homogeneous corpora. The system then transforms the template into a query which is applied to any number of frameworks. Queries return pointers to actual locations in frameworks that match the template structure. These units can be utilized immediately by analysis applications with minimal database related impedance mismatch.

A user can define a template structure with a graphical editor where the properties and relationships between



**Figure 2:** A representation framework with four domains: orthographic, linguistic, phonetic and acoustic. Units on different levels of each domain are used to model phones, phonemes, words, sentences, etc.

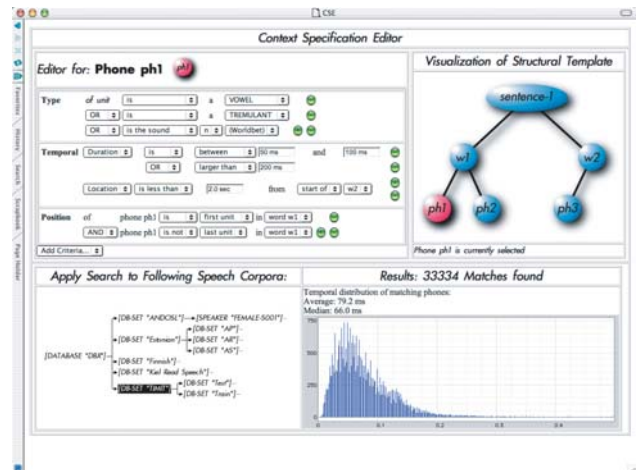
template units can be explicitly specified. In this project we are currently designing a *Context Specification Editor* (CSE) for this purpose. The creation of a template structure is depicted in figure 3. Due to the efficiency of the OO structure, an entire corpus can be queried for a particular context on the order of seconds. Therefore, while the user specifies the template structure, the system is able to simultaneously query entire corpora and indicate template selectivity by providing real-time results. Identical queries performed on a conventional RDBMS may take several orders of magnitude longer [6].

## 5.2 NETWORK BASED QUERIES

Efficient processing of large corpora requires systems with substantial working memory. Therefore, it may be desirable to locate these frameworks on a central database server with suitable resources. Another benefit of centralizing the frameworks is access. For example, a central repository can serve as a library of frameworks, queries and performed analyses that can be shared, and can act as a secure distribution service. Network latency can be reduced by having locally cached versions of the same raw data available on the user's machine. This allows the results of queries to be applied to data analysis applications immediately since no bulk data transfers are necessary.

## CONCLUSIONS

A project is currently underway in Finland that aims to improve the availability and use of shared multimodal speech corpora between different speech research groups. A conceptual map, which covers the relationships between typical speech entities in the form of metadata, is being developed and represented in RDF/XML. This map will form an integral core for applications such as speech



**Figure 3:** Database queries can be formed intuitively with a context specification editor. The selectivity of a structure can be shown statistically while the user develops the query.

collection systems, annotation editors, and database query systems. Four prototypical databases are being developed that utilize and comply with the above-mentioned methodologies.

## REFERENCES

- [1] *International Standards for Language Engineering*, <http://lingue.ilc.pi.cnr.it/EAGLES96/isle/>
- [2] *Corpus Encoding Standard for XML*, <http://www.xml-ces.org/>
- [3] O. Lassila and R. R. Swick, Eds., *Resource Description Framework (RDF) Model and Syntax Specification*, <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>, 1999.
- [4] K. Sjölander, "Recent developments regarding the Wavesurfer speech tool", *Speech, Music and Hearing Quarterly Progress and Status Reports*, vol. 44, 2002.
- [5] T. Altosaar, *Object-based Modelling for Representing and Processing Speech Corpora*. Report no. 63 / Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing. Espoo, Finland, 2001.
- [6] T. Altosaar, B. Millar, and M. Vainio, "Relational vs. Object-Oriented Models for Representing Speech: A Comparison Using ANDOSL Data," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH'99)*, vol. 2, pp. 915-918, Budapest, Hungary, September 5-9, 1999.