Three-Dimensional Modelling of Speech Corpora: Added Value through Visualisation

Toomas Altosaar[1], *Matti Karjalainen*[1], *Martti Vainio*[2]

[1]Acoustics Laboratory, Helsinki University of Technology, Espoo, Finland
[2]Department of Phonetics, University of Helsinki, Finland
Toomas.Altosaar@hut.fi, Matti.Karjalainen@hut.fi, Martti.Vainio@helsinki.fi

Abstract

Collections of annotated spoken language have formed an important basis for the development of speech technology. Their existence has promoted speech analysis research as well as enabled robust synthesis and recognition methods to be developed. However, many complex relationships remain unspecified within a corpus due to a lack of meta-data that describes the raw information in sufficient detail as well as the inter-relationships between signals, recording conditions, talkers, etc. A deficit of standards and formats, needed to express complex relationships, has also hindered the potential use and value of available corpora. This paper presents a novel three-dimensional model for exploring temporal as well as atemporal information existing in speech corpora. Examined are the potential benefits that are gained through corpus visualisation during the phases of creation, editing, verification, use, and exploration. The paper suggests that by providing a three-dimensional model of speech data, more of the inherent and potential value of a corpus can be utilised.

1. Introduction

A majority of speech applications — ranging from speech analysis through synthesis and recognition — require large amounts of data for analysis or training purposes. Speech corpora have satisfied this need by supplying an invaluable foundation onto which a robust understanding of the phenomena of spoken language can be gained. While the field of speech technology continues to become more refined, an increasing need for more accurate models of spoken language is required. Unfortunately, existing speech corpora often do not include the level of detail called for by advanced methods, or if they do, the relationships are not explicitly specified and it remains up to the user to form the missing links.

Although relationship-defining languages exist that can readily be applied to describing speech and speech corpora, e.g., XML [1], the relationships themselves need to be defined first. Simply converting an existing corpus into a mark-up language representation — without explicitly specifying any of the inferred relationships — offers minimal added value to users. To gain access to the full potential of a corpus, the relationships between objects must be explicitly specified and made readily accessible. Unspecified relations, as exist in currently available speech corpora, significantly reduce their potential value and ease of use.

This paper addresses the above problem by presenting a rich, accurate, and extendible model for speech and speech corpora that can be viewed and explored visually. The underlying model not only encompasses the familiar temporal aspects of speech, e.g., a signal and its related temporal annotations, but atemporal aspects as well. It can be hypothesized that the latter will play an ever increasing role in accurate modelling of speech, e.g., talkers, recording environment, equipment, etc., as more detailed and exact methods are developed for the science.

The three-dimensional visual environment can help the corpus designer define links within the corpus that would otherwise remain unspecified due to the complex nature of the task. Users of the corpus can benefit from a visual model of the data they plan to utilise and can thus understand and apply the data more efficiently and quickly to the problem at hand.

The visual environment developed in this paper relies on a detailed approach to modelling speech corpora developed within the QuickSig object-oriented signal processing system [2]. Here the object-oriented modelling paradigm is applied extensively so that information within a corpus is logically bundled into objects, e.g., signals, annotations, and talkers, which are given methods to interact with other objects. Also, explicit links are defined between objects, and besides giving objects knowledge of their surrounding environment, allow users to perform highly efficient database searches on the corpora [3]. Finally, a summary of the potential use of visualisation during the different phases of corpus creation and use is presented.

2. Object Modelling of Corpora

Visualisation can be seen as a mapping from an object model to a visual space. This however requires that the object model must first exist. In QuickSig the object model for a specific corpus consists of a large set of instances created from an extensive library of classes modelling speech phenomena. QuickSig has been applied to many different speech related applications [4]. It is written in Lisp and CLOS, making it a dynamic and reliable explorative research platform.

2.1. Object Model for Temporal Data

In QuickSig data from a corpus have their type information reinstated through the use of Corpus Specification Models (CSM). CSMs, defined for each different corpus under study, e.g., TIMIT, Kiel, ANDOSL, etc., contain meta-data describing information that is missing or assumed, which must be otherwise provided by the corpus user. For example, a CSM supplies for some specific corpus the signal waveform format, character set,

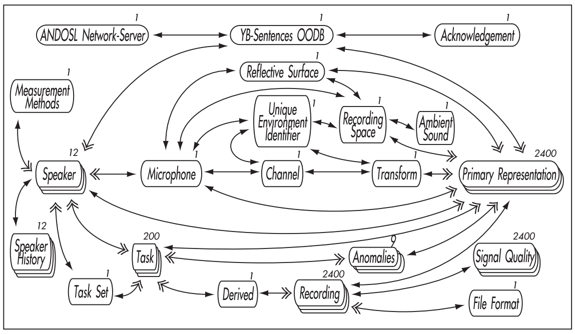

Figure 1. A flat, two-dimensional object-layout map of ANDOSL "YB Sentences" data. This object-oriented model of the atemporal data is built using 21 classes.

phonetic alphabet, labelling syntax used in annotations, and corpus structure.

Once raw data from a corpus have been tagged, a generic compiler and linker void of domain specific knowledge are applied to form a representation framework where data from different spoken language theories, e.g., linguistics, orthography, and prosody, can be represented concurrently. Each utterance in a corpus is transformed into a framework consisting of:

Domains:	e.g., acoustic, linguistic, phonetic, orthographic, prosodic, etc.
Levels:	e.g., segment, phoneme, syllable, word, sentence, etc.
Units:	e.g., actual segments, phonemes, phones, words, etc.

Database access is performed on the frameworks by structural matching. For example, if a certain sequence of phones is being searched for, a predicate function is generated that detects the desired context. The function is applied then to every phone in every framework over which the search is to be performed. Contextual matches return the actual objects instead of just weak descriptions, and can be used immediately in further processing. Searches are efficient due to the existence of links (both explicit and computational) and usually no indexing of data is required, as is common in relational-database management systems (RDBMS). Frameworks can be stored to persistent memory through the object-oriented database management system (OODBMS) available in QuickSig.

2.2. Object Model for Atemporal Data

In a similar way to temporal data, an object model is first developed for atemporal data that captures the non-temporal features of a corpus. Starting from the mono- or dialogue of the speaker(s), and including the set-up of the recording environment (e.g., microphones used, physical location of talkers and microphones, equipment layout, room acoustics, etc.), as well as information about the talker(s) such as their personal history which affects their accent, a collection of objects describing the recording is developed.

For example, the ANDOSL [5] corpus was selected as an exemplar collection of atemporal speech due to its rich description of the recording environment. Using an object modelling approach, data available within ANDOSL can be mapped out into a flat two-dimensional object map that logically indicates what information is available in the corpus. This is seen in Fig. 1 where data from the ANDOSL YB_Sentences is separated into 21 different object classes.

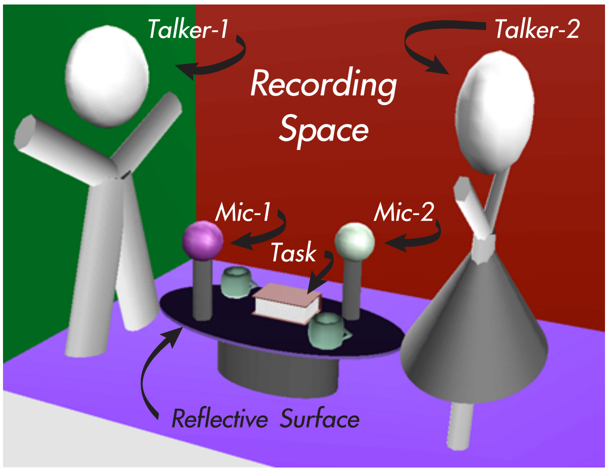

Figure 2. Modelling corpus objects in a recording space: talkers, microphones, a task set (dialogue instructions), and physical objects in the room such as a reflective surface. Objects are aware of their relationships to other objects through links.

3. Three-Dimensional Visualisation

Once the object models have been formed the mapping to a visualisation space can be made. Since the temporal and atemporal models of speech within QuickSig are separate but linked entities within the same integrated computational environment, two different visualisations have been created.

3.1. Atemporal Visualisation

Each object class is first assigned a visual token, e.g., a talker is represented as a simplification consisting of a head, two arms, and two legs. Each instance of a class can be visually modified according to its specific internal values. For example, if the talker is of the female gender, an extra cone is drawn around the vertical axis depicting a skirt, as seen in Fig. 2. Other objects, such as the recording environment, placement of the microphone(s), the task of the talker(s), and items such as reflective surfaces, etc., can be included in the visualisation. A three-dimensional atemporal model of two speakers participating in a dialogue can be seen in Fig. 2. Of significant importance is that there exists a symmetric mapping between the underlying object model and its visualisation(s). This allows for user interaction enabling editing to be realised.

Once visual tokens have been assigned to each different object class existing within an object model, the entire atemporal data space can be visually rendered. Figure 3 shows a perspective of ANDSOL atemporal information from two different corpora within the 30 available ANDOSL CD set: "YB Sentences" and "MAP TASKS MMK_FMK1". Since the two different corpora are merged in the underlying model (e.g., a single object represents a common talker in both corpora), no duplication of identical objects takes place. The user is therefore presented with a logically true model of the existing information.

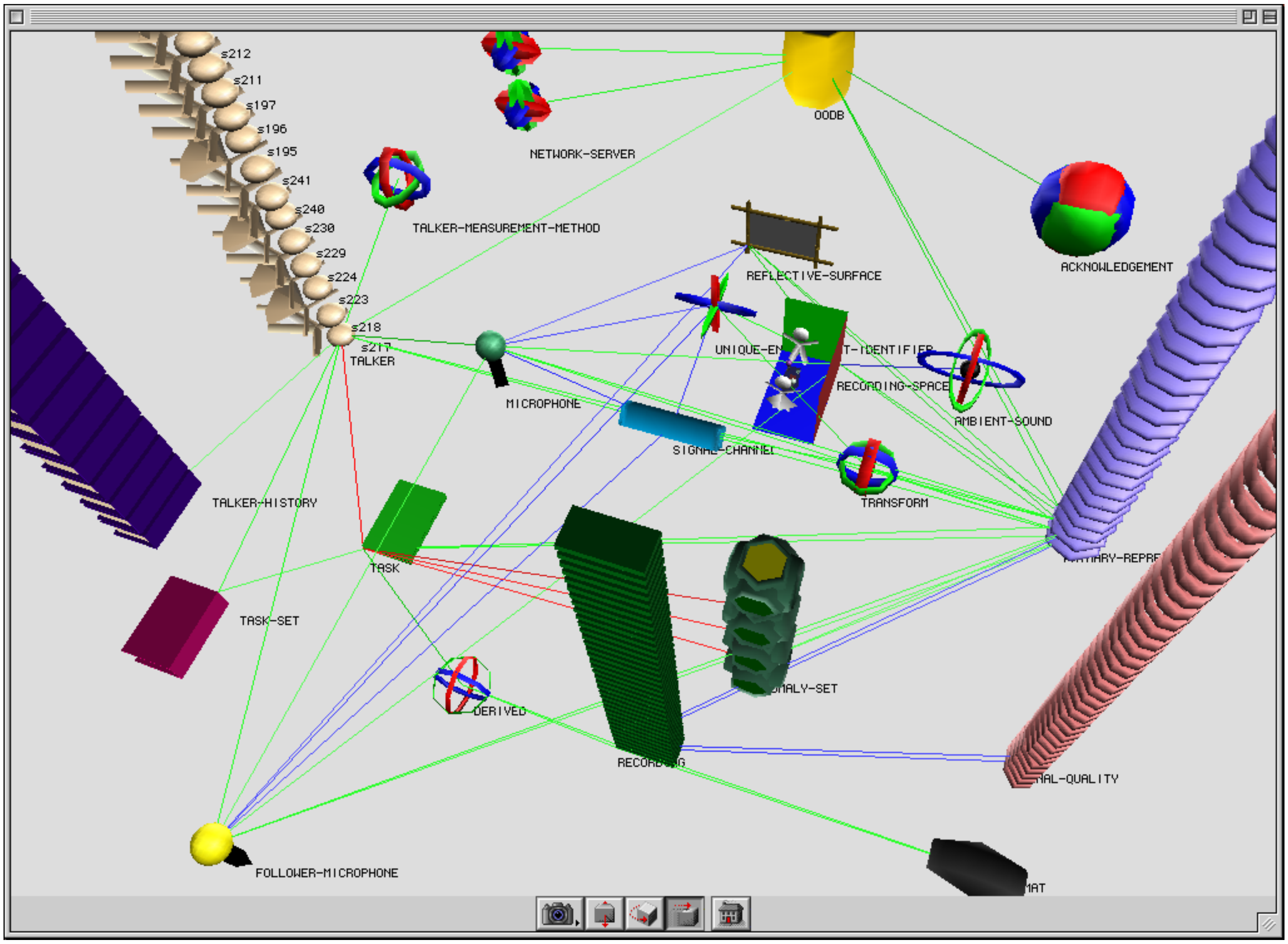

Figure 3. Visualisation of ANDOSL atemporal data in three-dimensions: the user can "fly" through the space with the aid of the mouse and explore/edit objects and their relationships. In this figure Talker S217 indicates his influence (direct relations) through 1-to-1, 1-to-many, and many-to-1 links to other objects in the space. The stack of RECORDING *objects, near centre-bottom, are the actual audio recordings, which include temporal annotations. One recording is shown in Fig. 4.*

Also seen in Fig. 3 are links emanating from speaker S217 to other objects. These indicate the direct interaction speaker S217 has had on other objects, such as the task set and task used (what he read), the microphones he spoke into, the electrical characteristics of the signal channel and any possibly applied transform, etc. Only a few tens of links are shown since talker S217 was requested to show his direct relations with the corpus' other objects. The underlying model for this two-corpus case has over 70.000 automatically generated links and visualising all of them at the same time would not be of practical use. Traversal from one object to another, that are not linked directly, is accomplished by "computational" links, i.e., method functions are automatically generated to allow reaching another object via the shortest linked path. For example, in the above figure it can be seen that talker S217's voice exists in only two of the recordings (two links exist between the *PRIMARY-REPRESENTATION* and *RECORDING* stack of objects). Many benefits exist in having objects knowledgeable about their location in the environment. For example, subsets of the original corpus can be easily formed given some criteria, e.g., "*Generate a subcorpus where only talker S217 exists*" (objects with links in Fig. 3 are then saved separately), or, "*Create a subcorpus where only a certain microphone was used*".

Finally, users can position themselves by "flying" to any desired location via mouse control. The user is then able to select an object and inspect/edit its values and make changes to the underlying object model.

3.2. Temporal Visualisation

Any of the *RECORDING* objects from the atemporal visualisation can be viewed in more detail along with its annotations. Figure 4 shows a recording from the Kiel Corpus of Read Speech [6] (Kiel data chosen since it is richly annotated) where the audio signal as well as orthographic, linguistic, and phonetic domains, levels, and units make up a linked structure. Again, the user is able to move around in the space and explore, e.g., listen to units audibly by clicking on them, look at unit properties, generate queries by selecting template structures (e.g., find all cases where an utterance initial word has three syllables, and the second phone uttered is a [a] (IPA), etc.

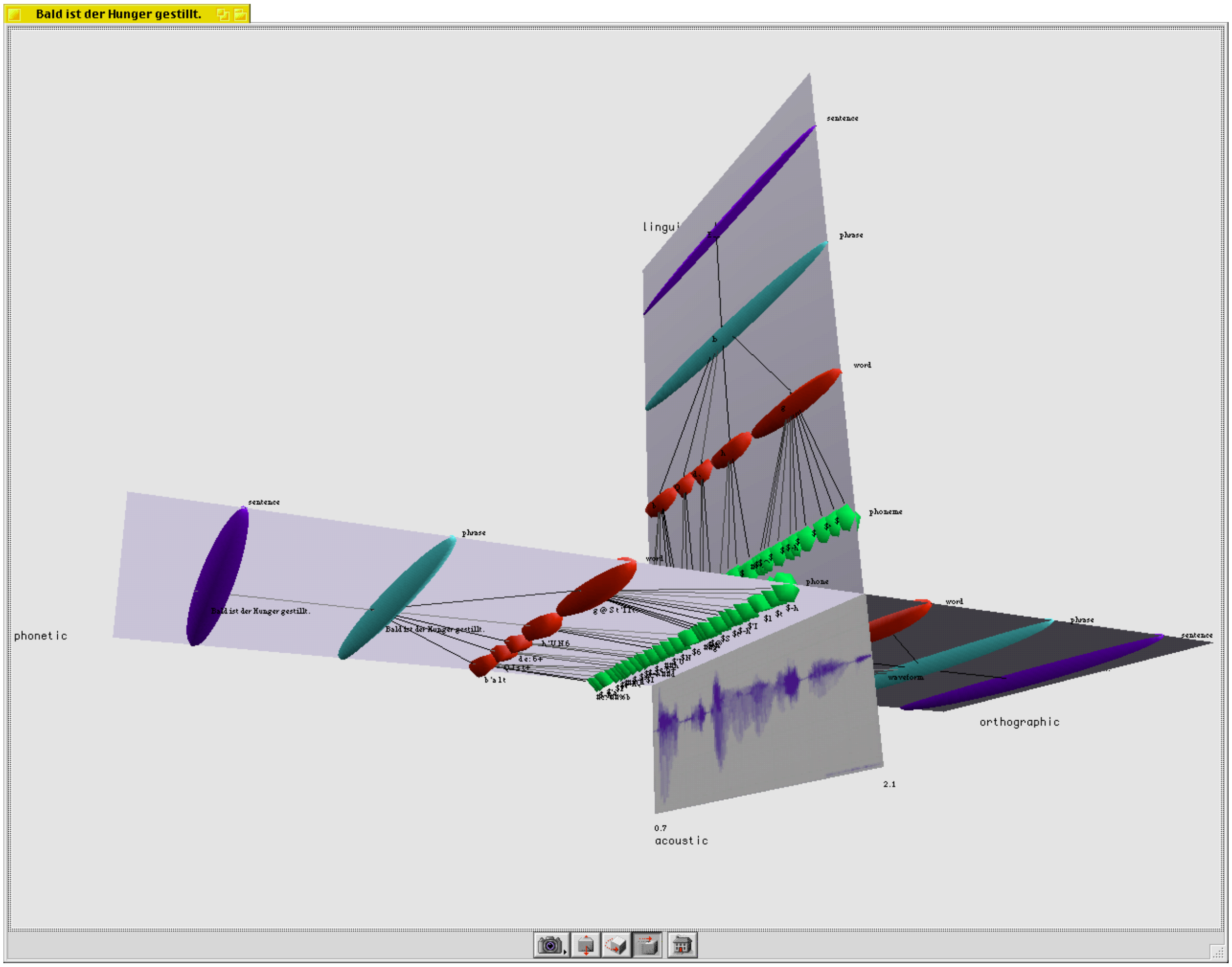

Figure 4. Temporal visualisation of the representation framework for the Kiel sentence "Bald ist der Hunger gestillt." From the bottom and working clockwise: the acoustic, phonetic, linguistic, and orthographic domains that are composed of separate levels consisting of speech units. Database queries operate on the units as well as the links between them. Inter-domain links exist between some units as well, e.g., between a phone and its related phoneme, and can be used to detect deletions, replacements, and insertions.

4. Potential Use of Visualisation

Some of the benefits that may be gained through corpus visualisation during different phases of corpus interaction are listed below. During corpus

Creation:	new objects can be visually created and added to a corpus,
Editing:	objects can have their internal values and relationships specified, manually if needed,
Verification:	removal of redundant information, automatic creation of meta-data and relationships
Use:	semi- or automatic query formation possible, select part of data for training,
Exploration:	ease investigation of unfamiliar corpus contents by a potential user.

5. Summary

Three-dimensional visualisation of collections of speech in the form of corpora, as well as speech signals themselves including their annotations, can provide an improved user interface to the data, than e.g., a flat two-dimensional representation. Visualisation can add significant value to a corpus and promote its use in speech analysis and application areas.

6. References

[1] World Wide Web Consortium (W3C). (2000). Extensible Markup Language (XML). URL: http://www.w3.org/XML/

[2] Karjalainen M., Altosaar, T. & Alku, P. (1988). QuickSig—An Object-Oriented Signal Processing Environment. In Proceedings of the 1988 IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 1682-1685. New York, USA.

[3] Altosaar, T., Millar, B. & Vainio, M. (1999). 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. Vol. 2, pp. 915-918. Budapest, Hungary.

[4] Altosaar, T., Karjalainen, M., Vainio, M. & Meister, E. (1998). Finnish and Estonian Speech Applications developed on an Object-Oriented Speech Processing and Database System. In workshop proceedings of First International Conference on Language Resources and Evaluation. Speech Database Development for Central and Eastern European Languages. Paper No. 6. Organised by the BABEL Project, Copernicus No. 1304. Granada, Spain.

[5] Millar, J.B., Vonwiller, J.P., Harrington, J.M., Dermody, P.J., "The Australian National Database Of Spoken Language", Proc. ICASSP-94, Adelaide, 19-22 April, Vol. 1, pp. 97-100, 1994.

[6] Simpson, A., Kohler, K., & Rettstadt, T. (eds.). (1997). The Kiel Corpus of Read/Spontaneous Speech - Acoustic data base, processing tools and analysis results. Arbeitsberichte (AIPUK) Nr. 32. Universität Kiel, Kiel, Germany.