ABSTRACT

In this paper we present a general framework, based on the Object-oriented paradigm, for modeling and designing a model of speech data representation, and we propose a particular use of it for cineradiographic data, including sagittal views of the vocal tract, frontal pictures of the lips, and acoustic signals. We introduced semantics to represent relationships between speech objects. Thus we adopted the concepts of primary data, that means either the raw data (recorded signals and images) or their related descriptive data (information on speakers, corpora and recording conditions), and of derived data, such as vocal-tract's contours, sagittal distances, area functions, or any other possible measurements taken from X-rays pictures. Indeed, the notion of derived data model has been useful for users, to manage raw data and results of data analysis in the same way.

1. INTRODUCTION

It is now well-known that the development of speech databases is of great importance to the progress of speech technology, and in particular for speech recognition purposes. The usefulness of databases for more fundamental research is less often acknowledged. However, the need for related acoustic and articulatory data has been clearly identified, to develop models of the vocal tract, and to understand its control during speech production.

The French speech community, essentially through the Institute of Phonetics in Strasbourg (IPS), successively directed by G. Straka and P. Simon, has been very active in recording a large amount of X-Ray films, on several speakers of various languages. In collaboration with IPS the “Institut de la Communication Parlée” (ICP) is currently designing a multimedia platform, including a database, to manage this kind of data, and to facilitate their accessibility and processing to the speech community. Three types of data are available: X-ray films for profile views, extracted from Bothorel et al., 1986[1] see Fig.1, at 50 frames per second, Labiofilms and Audio signals.

The main goal of this project is to develop a software environment making easier the access to these data as well as their analysis and their use for further study on speech production.

Including X-ray images as raw data in the database is considered in the database schema. These data increase the complexity of speech data structure and give a multimedia dimension to the data modeling and their management. Hence, multimedia management functionality such as computer-based viewing of X-ray and vocal tract image sequences is necessary. In this perspective we have proposed a methodology for the data organization and for a future storage of this Multilingual X-ray database on CD-ROM.

To reach the objectives of this project, we based our approach on Information Systems (IS) principles. We introduce the term SpeechRIS (Speech Research Information System) to mean a system that should be able to incorporate one or more speech databases, to deal with complex data structures and their dynamic evolution as well as to enable an efficient storage of large amounts of data. To improve our chances of success, we have followed an approach that takes into account the nature of the data, and then designs the system from a user's perspective. In addition, the feasibility of the designed system is assessed.

As regards building SpeechRIS, similar questions faced by other scientific IS should be adressed:

1. What is the appropriate data model?
2. What is the appropriate database architecture?
3. What minimum technological support is necessary to achieve display of graphics, images and video?

1 Now at CLIPS/IMAG
2. MODELING PRIMARY DATA

Data handled by speech researchers consist of heterogeneous and strongly inter-connected entities: physical signals, articulatory data, various kinds of labels and parameters, information on speakers, signal processing programs, etc. The increasing volume and types of data require appropriate management solutions[2][3][4].

The major drawbacks of the relational model are the limitation on data-structuring flexibility and the lack of ability to explicitly specify semantic information about relationships. Object-oriented modeling is well suited to adequately represent the relevant complex relationships between the various entities in the real world. In addition, object modeling allows to combine data and their processing at the same level [5].

To help in data representation for speech databases, at ICP, [6][7] a formalism has been proposed that can be summarized as follows: A Speaker S pronounces a Corpus C and produces a Realization R that contains a collection of Signal Elements SE in a given Recording Condition RC. Thus a SpeechDatabase is a set of Realizations. The formalism has been implemented in the SIDOC system developed at ICP. To make use of this concept, a Realization class has been defined that is related to three other classes describing the experimental protocol in a recording session: a Speaker class, a Corpus class and a RecordingCondition class (Fig.2). Some of these classes are defined as collections. The Corpus class contains information about tokens and is structurally defined as a collection of linguistic corpus elements to be recorded, such as words, sentences, numbers, etc These tokens are represented in the CorpusElement class which points to a Transcription class, which contains the orthographic and phonetic transcriptions of the token.

VocalSignalElement class is related to the portion of signal, that corresponds to the class ElementRealization, and is described in the CorpusElement class. This class has two instance variables, which are integer values: BeginAddress and EndAddress. This class allows retrieval of the portion of a signal associated with the ElementRealization, without the need to edit the entire signal (Fig2). Name is the only attribute of the class File. This class points to the class Medium to retrieve any file that may be stored on a mass storage support such as CDROM.

Medium class describes the Type of the mass storage device (CDROM, DON,...) and the Name of the concerned Speech Database, containing the desired file.

The extension of the data model for the representation of different kinds of data (X-ray and Video films) includes two abstract classes X-rayFilmRealization and VideoFilmRealization that inherit all information from the topmost class Realization. This model allows us to keep trace of all descriptive data in the case of films realizations. The model is depicted in Fig.2 and Fig.3.

We have used the Gemstone system (a fully Object Oriented Database Management System) as a basic layer to implement the model. A flexible database schema has been then designed.
3. MODELING DERIVED DATA

Different researchers may apply different algorithms to reach different objectives. In order to make use of the results or data obtained by other scientists and to allow researchers to apply new processing on the same data, we must have a full understanding of the data derivation history - how they were produced (from which base data? with which parameters? and which actors?).

3.1 General Principles

At the upper level, derived data are modeled in SIDOC by the class DerivedObject. This Class is the topmost class in the derived data model. As such, its superclass is Object class (Root of hierarchy of tree). There are no class variables, but two instance variables identify each instance (e.g each specific derived object): creator and baseObject. We distinguish two kinds of Agents: Human Experts and Softwares (e.g Expert Systems and programs) (Fig.4).

![Fig.4. Speech Research Actors](image)

PeriodicalEstimation refers to data that are directly extracted from raw data, such as formant value, and sagittal contours. Thus we distinguish two subclasses Formant and VocalTract_Contour. An instance variable samplingRate was included to indicate the interval of time at which these data are sampled. Label class describe discrete occurrences of events at particular places in the signal.

At a second level we have represented the derived data extracted from the periodical estimations such as lip height, sagittal dimension,... We distinguish four classes ProfileMeasurement, FrontLipMeasurement, SagittalFunction and AreaFunction. Each of these classes has its own instance variables, which are integer values, and it refers to the file where related data may be retrieved.

The link between derived data and the primary data represented in Fig.4 is made via an association of derived data with the class ElementRealization. This modeling allows the interface to interact directly with derived data and retrieve immediately their corresponding primary data, such as signal and X-ray image file names and other information.

3.2 Modeling Representation of VT geometry

Queries to the database world aimed primarily at finding shapes of vocal tract which correspond to the desired segments. Example: displaying contours associated to the phoneme /g/ occurred in the sequences /aga, agu, agi, ago, age, agy,.../ for comparison.

We conclude that the processing of queries has to handle several aspects: a VT contour can be found, on the one hand, according to the semantics of the objects composing the contours, and on the other hand according to the derivation semantics. The contour model has to take into account the evolution of contour structure. Each contour is therefore described structurally (as a composition of objects which are sub-contours) and geometrically (as a sub-contour defined as a set of points), corresponding to its spatial view. Our aim is to allow users making queries directly on the different structures of the vocal tract, a specific class has thus been introduced: VocalTract_Contour. This class has two attributes, the Name and the shape as a set of sub-contours (Fig.6). Hence this model allows users to represent an infinite number of sub-contours which may be updated.

![Fig.5 Derived data model](image)

![Fig.6 Vocal tract structure](image)

4. SIDOC ARCHITECTURE

If a database management system is to be truly useful in the speech research environment, much care must be taken of how it will be used and, in particular, how it will be interfaced to the outside world. Our approach consists in querying descriptive data related to various realizations depending on the speaker, or on the speaking conditions, or on the linguistic context, as the first step to any session intended to extract a large number of tokens.

The second step of such a session consists in querying related labels and respectively derived data stored in the database system, and exporting the results for further processing by softwares reachable from the SIDOC environment (i.e. MATLAB, or ISIS[8]).
5. Query Interface

The query interface cooperates with the signal processing software (ISIS)[8] developed at ICP. Thus, queries are processed by the SIDOC interface (Fig.8), but the visual synchronisation of acoustical signals and X-ray images with their derived data is made by the functionalities of ISIS system. An interface allows the users to make the desirable query: Fig 9 displays an example of output query, including both raw data and derived data. On the left panel the signal is displayed, the middle shows the derived sonagram, and the right panel presents the vocal tract contour corresponding to the time specified by the cursor placed on the signal as shown.

CONCLUSION

Our investigation consisted in developing a data model to represent basic speech data entities used mostly in the speech research community. This model has been useful in managing all speech databases available at ICP. We have extended the data model to integrate Cineradiographic data (X-ray films as primary data and their derived data). The data abstraction techniques such as inheritance were found to be essential for describing either primary data (X-ray realizations and Video realizations) or derived data and their components.

ACKNOWLEDGMENTS

We are very much in debt to A.BOThOREL, P.SIMON, F.WIOLAND, J.P.ZERLING & R.SOCK from IPS & T.LALLOUACHE who kindly made available to us the data. The authors would like to thank G.BAILLY and P.BADIN for their critical comments and suggestions.

REFERENCES