MOOSE: Management Of Otago Speech Environment

Laws, M. and Kilgour, R.

Knowledge Engineering Lab
Department of Information Science
University of Otago
New Zealand.
email: maaka@kel.otago.ac.nz

ABSTRACT

With the advent of spoken language computer interface systems, the storage and management of speech corpora is becoming more of an issue in the development of such systems. Until recently, even large corpora were stored as individual text and speech files (for example, [18]), or as a single, monolithic file (for example, [1]). The issues involved in management and retrieval of the data have been, to a large extent, overlooked. Initially, the Otago Speech Corpus files were stored in a nested directory structure, including speech and text. Although each item of data could be retrieved, there was no relationship between them. Therefore, a phone extracted from a spoken word had no direct relationship back to that word. Furthermore, with no management system, a new user of the corpus was required to become familiar with a cryptic naming scheme and the nested directory structure before the corpus was useful. Relational database management systems (RDBMS) are proposed as an ideal tool for the management of speech corpora [5]. Relationships between words and phonemes, and the realisations of these, can be stored and retrieved efficiently. RDBMS may be constructed with various levels, to store speaker, language, label transcription, and phonetic information, plus speech as isolated words, and derived segmented units. In this manner, the data becomes a complete speech corpus. An implementation of such a system is presented to manage the Otago Speech Corpus, currently called Management Of Otago Speech Environment (MOOSE). The ability of the MOOSE to be applied to other corpora is currently under investigation.

1. INTRODUCTION

1.1 Database Design

Databases can range from simple, single files, to very complex, multiple-file, distributed client-server systems. Within this range, there is a requirement that each database be an effective robust tool for the user of such a system. Database costs are also associated with size, where the larger the database, the more expense required to create, document, distribute, and maintain the system. Database design can be a simple or complex process, depending on the scope of available resources to the developer. Current personal computing trends call for the design of databases that allow users to access a vast electronic storage of facts. Data stores that can be in a multitude of different formats represents these facts. Data store formats can be; plain characters, formatted text, numbers, logical codes, symbols, tables, figures, graphs, pictures, photographs, graphics, sounds, speech, animations, video, movies, and music—all can be collectively termed as electronic files.

All database computer programs keep track of the electronic files. Moreover, database designs can vary considerably from one program to another. However, all designs are usually based on a similar structure that is driven by the actual user's requirements. Meeting those requirements is closely dependent on the types of electronic files used [5, 13].

1.2 Speech and Language Databases

For many years now, linguists have required a variety of database structures to store and manage their speech and language data. Linguistic databases can facilitate a host of application related uses, from data collection, storage and retrieval to the in-depth analysis of the speech and writing systems. The following have reported using speech and language databases to assist with their linguistic research: [2, 3, 4, 6, 7, 9, 10, 14, 16, 18].

Many of these databases use a ‘flat-file’ structure to save the data stores, using a complex system of indexing. The data is usually a large single coded file in a directory with other specialised indexed files [10, 15]. File-based structures are reliable systems if the data store is small, with simple formats and basic user requirements, but as their size and functionality increases, so do the complexities associated with managing the data and index files. There is a point when the data stores become too large, that the database applications perform as inefficient processors, with little or no more flexibility for speed, expansion and adaptation [15].

1.3 Relational Databases

The relational database is now the most preferred conventional software engineering technique for the management, access and retrieval of large electronic files or data stores. When application interfaces are linked to relational databases using the Structured Query Language commands (SQL), the data is more effectively transformed into valid, meaningful, and timely information. The user can instantly interact with the database, to either gain knowledge, make decisions, perform tasks, control devices, or even be entertained [5, 8, 9, 13, 15].

Sinclair and Watson [15] have reported on the initial development of the Otago Speech Database that was first modelled on the Relational DataBase Management System (RDBMS). The motivation was to give the user the ability to query (using SQL) all the data types for better functionality, and also for future expansion. The current Otago Speech RDBMS developments now provide a full and comprehensive database system with the required speech and language information designed specifically for speech recognition. It contains the speakers’ and recording details, English and Maori words, sentences, phonetic transcriptions and pronunciations, segment labels, and a growing set of digitised Maori and English speech examples of words and phonemes.

2. DEVELOPMENT OF MOOSE

2.1. Existing Otago Database structure

The initial phase of the HySpeech project [7] was to have a large enough set of speech data that could be used to accomplish the growing demand for analysis, training, testing and validating of the various modules associated with the recognition model [15]. The collection of the New Zealand English speech data was undertaken in a rigorous systematic format to bias the data with an inherent New Zealand accent. Currently there are thirty-six New Zealand English speakers, 9420 uttered words, and 8123 segmented phonemes. The database was constructed with various levels to store speaker, language, label transcription, and phonetic information, plus speech as isolated words, and derived segmented units - a complete speech corpus.

The ongoing development of the Maori speech database is based on the NZ English corpus [15], where new and existing Maori speakers and words will have similar file formats and codes [11]. There are currently seventeen speakers of Maori in the corpus. Presently a small selection from one male speaker's speech data has been hand labelled and segmented into diphones (a phonemic transcription of two phones). This was for the intended purpose to build an experimental diphone unit concatenative Maori text-to-speech synthesiser, to form part of HySpeech/2's Bilingual Speech Interface [9].

Other work has concentrated on developing the English and Maori lexical structure, to provide a medium sized text database for simple headword translations [2]. The database has 3000 English words with phonetic transcriptions (2000 basic words with IPA codes and 1000 scientific words), and over 4,700 Maori translations available, on average there are two Maori words to each English word [17].

2.2 The Relational Model

To build a RDBMS, a structured approach must first be considered to form the basic building blocks that represent all relationships between the previously mentioned data formats. An Entity can be anything known as a fact that is in a logical form. These facts are classed as Data. For example, in the Otago speech database, the 'English Words' are entities. Each entity must also be identified by certain characteristics that make it part of a larger set, these are called Attributes. The attributes for the entity 'English Words' are 'WordCode' and 'EnglishWord', where the first attribute is used for indexing purposes, and the other to list all the words. Collections of these related entities are commonly known as an Entity Set. Relationships between entity sets are based on how entities interact with other entities within other entity sets. For example, the relationship between the Otago speech database entity for 'English Words' and the 'Maori Words' entity are 'Translate'. This relationship is best described in a more meaningful form by a further entity called 'English-to-Maori. To complete the database structure, there are three classes of relationships used, they are One-to-One, One-to-Many, and, Many-to-Many [13]**.** The Otago speech database 'English-to-Maori' entity uses the one-to-many relationship when one English word relates to many Maori translations.

A Relational Model can be constructed to represent a solid theoretical layout of a proposed database structure. Relational models have a more conceptual approach to the design of the database, and are therefore very simple to understand once the terminology and the Entity-Relationship Model (or E-R Model) symbolism's are known [13]. Tables are the main relational construct that combines both entities and attributes to form entity sets. All tables have the attributes marked in columns and entities layered in rows.

A schema (or detailed report) can be generated to provide a detailed description of the database tables and 'Query Structures'. The following E-R Model was designed for part of the HySpeech database [11, 15] with the resulting schema used as the basic structure when developing MOOSE. Data and their relationships, such as Primary and Foreign Keys, are represented in the RDBMS. This allows the user to query or validate the structure and the rules therein.

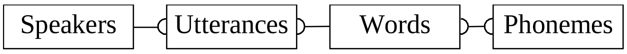

Figure 1.1 Simplified E-R model of the MOOSE system.

2.3 New Otago Speech RDBMS

There are four main areas that form the basis of the RDBMS. The English and Maori words and phonemes are in separate tables, for example, see Tables 2.1 and 2.2. The 'Speaker Background Data' currently holds data such as that in Table 2.3. The 'Speech Data' and 'Segmented Data' contain the WAV files. The 'Speech Corpus' is the amalgam of these tables. The database engine is the core of the entire system - it holds the tables, queries, metadata, relationships, and the interface modules.

2.3.1 New and Existing Tables

The Otago speech RDBMS provides the formal file based structure for the English and Maori word lists and the selected speech files are stored and managed. The prime function of these newly developed English to Maori tables has been incorporated to provide the 'Lookup' or 'Find' SQL functions with the interface. An English word will be retrieved along with its Maori translation(s) and embedded speech files, the databases SQL 'Select' command will display the words and the 'Object Linking and Embedding' (OLE) function will play the file automatically upon launching the appropriate sound application (for example. SoundRecord).

2.3.2 Other Tables

The new information on speakers and phoneme labels are now maintained by the RDBMS, Tables 2.1, 2.2 and 2.3 outline the type of data statistics that can be produced.

Symbol	Count	#Speakers
p	83	8

Symbol	Count	#Speakers
b	61	8
t	196	25
...	...	
E@	37	8
q	1	0

Table 2.1 Sample of the phoneme segmentation list, showing the number of segmented phonemes extracted from the corpus.

Spkr	Word Code	Utt.	PhCode	Start	End
02	001	1	001	00002422	00003092
02	001	2	001	00002205	00003058
02	001	3	001	00003629	00004752
02	002	1	002	00008457	00009536
...	...	...	...	...	...

Table 2.2 Examples of the phoneme labelling information that is currently housed in the database for automatic segmentation by the 'off-line' Signal Processing Module

Speaker	Sex	Age	Language	Dialect	Education
01	M	20	English	NZE	University
02	M	21	English	NZE	University
03	F	16	English	NZE	High School
...	...	...	...	...	...

Table 2.3 Section of the speaker information list.

3. ENGLISH AND MAORI DATABASE MODULES

An advantage of using SQL is that the system becomes more functional due to the availability of features for queries, updates, and reports. This allows access to a bilingual dictionary of the two languages, even if they are stored as separate entities.

3.1 English-Maori Text and Speech Query Table

Table 3.1 contains combined examples of both the English and Maori words, and the Maori speech files, generated using the SQL 'Select' command. In the table, a few examples of exported TEXT from the 'English-to-Maori Select Query Table' are shown. This option provides the user with an assortment of tools to present the data in a more meaningful way that can answer 'spur-of-the-moment questions' [13].

WodeCode	English	Maori	Speech
1202	of	a	a.WAV
92	and	ma; hoki; na; me	aa.WAV
2002	yes	ae	aae.WAV
...	...	...	...
600	eat	kai; kainga	kai.WAV
1177	no	kore; kao; kahore	kaaore.WAV
76	all	katoa	katoa.WAV
...	...	...	...
1985	woman	kui; wahine	wahine.WAV
1930	water	wai	wai.WAV
901	house	whare; whare kainga	whare.WAV

Table 3.1 An example list of the text and speech query, sorted on 'Maori Speech', which shows all the speech files that are linked to the database table

3.2 English-Maori Text

Another version of the database query table that only contains the English and Maori words as text can be generated. This could also be used with the database where the same search and extraction are used, but the digitised speech examples are stored as individual WAV files in a separate (but designated) directory. They are also accessed and played through the BSI in the same way. This method is ideal for keeping the overall Maori speech table storage down to a minimal size, with the added advantage that the WAV files can be easily upgraded without having to delete, add, or embed new sound files into the existing table. This can be done separately at any time. This method could also be used, if and when the speech examples are pressed onto a CD-ROM for distributed access.

4. OTHER DATABASE FEATURES

Presently the text and speech data exist in the form of 'crisp records' contained in conventional look-up tables, linked to word associated digitised speech examples. The database is progressively extended to allow for more English and Maori words with many variations in their meaning and uses, plus more speech examples in both languages. For example, English Male/Female and Maori Male/Female, or even 'Fuzzy Data Sets'. Thus, a 'Similarity-Based Query Module' [7] could be developed to allow the user to perform fuzzy-like ambiguous queries using 'Fuzzy Driven' SQL commands on the RDBMS.

Furthermore, the system has the potential to be expanded to included gender recognition. This would identify the relative pitch of the speaker's first and second formant frequencies. A gender identifier (for example. M or F) would be used to select and play the appropriate female or male speech WAV file, from either a binary field or a directory.

Multimedia interfaces to databases have become the next clear shift in demand by users requiring multi-modal forms of communication and interaction [8, 9]. With this demand, there has come the expansion and extension of I/O peripherals available to control such forms, currently ranging from speech recognition and synthesis to the more sophisticated virtual reality (VR) headsets and gloves [12].

5. SUMMARY

The current implementation of an intelligent human-computer interface to control a specialised speech and language database means that a change in computer interaction from a manual mouse and keyboard mode to an automatic speech control mode is facilitated between the user and the database. This change will be emphasised by the further need to provide more comprehensive data sets that can fulfil future demands for a complete move away from the manual control devices to automatic speech understanding and then onto total VR interaction [12].

The current database structure allows for multi purpose functionality. Queries can be used to retrieve whole words, segments of words and speaker data. The integrated design allows for little or all of this information to be entered and subsequently used.

By explicitly defining the relations between the data, two-way translations can be performed. Thus the English-to-Maori relations relate to the Maori-to-English translations. This extends the functionality of the database to a bilingual system.

Whereas previous implementations stored word data as well as word-segment (phoneme) data, the RDBMS allows the definition of additional segments, such as diphones. Additionally, relating the segment to the existing words eliminates redundant storage of explicit word-segments. Queries for phonemes can thus return phoneme data automatically generated from the words in the database.

The automatic segmentation is performed by externally defined functions. There is the potential for the inclusion of additional functions to the system. In this way, a more integrated environment is created for recording, segmenting and otherwise managing the maintenance and development of speech corpora.

Currently, the ability of the MOOSE system to generalise to other corpora is under investigation. The relations have been defined to allow for the storage of sentences, and for the storage of continuous and spontaneous speech. Speaker information includes language and dialect. Therefore, corpora such as TIMIT [18] may be stored using the same management structure.

Acknowledgement: The authors would like to acknowledge the assistance of PGSF grant UOO606, supported by the NZ FRST. Further information can be found on the World Wide Web at "http://kel.otago.ac.nz/CBIS/". Thanks also to Mike Watts for his input into the ER Model.

REFERENCES

1 Bauer, L. (1994) *Introducing the Wellington Corpus of Written New Zealand English*, Te Reo, Journal of the Linguistic Society of New Zealand, Volume 37, University of Auckland.

2 Benton, R. A., Tumoana, H., Robb, A. (1982) *Ko Ngä Kupu Pü Noa o Te Reo Maori: The First Basic Maori Word List*, New Zealand Council for Educational Research, Wellington.

3 Cole, R., Noel, M., Burnett, D. C., Fanty, M., Lander, T., Oshika, B., Sutton, S. (1994) *Corpus Development Activities at the Center for Spoken Language Understanding*, In Proceedings of the ARPA Workshop on Human Language Technology.

4 Crystal, D. (1992) *The Cambridge Encyclopedia of Language*, Cambridge University Press, Cambridge.

5 Date, C. J. (1990) *An Introduction to Database Systems*. Volume 1, 5th Edition. Addison-Wesley Publishing Co.

6 Hunt, A. J. and Black, A. W. (1996) *Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database*. Proceedings: ICASSP, Atlanta, GA.

7 Kasabov, N., Sinclair, S., Kilgour, R., Watson, C., Laws, M., Kassabova, D. (1995) *Intelligent Human Computer Interfaces and the Case Study of Building English-to-Maori Talking Dictionary*, Kasabov, N., Coghil, G. (Eds.), Proceedings ANNES 95' University of Otago, Los Alamitos, CA: IEEE Computer Society Press, pp. 294-297.

8 Kasabov, N., Kozma, R., Kilgour, R., Laws, M., Taylor, J., Watts, M., Gray, A. (1997) *Neuro-fuzzy Techniques for Speech Data Analysis and Adaptive Speech Recognition*, In: N. Kasabov and R. Kozma (eds) Neuro-fuzzy Tools and Techniques, Physica Verlag, Heidelberg

9 Kasabov, N., Kozma, R., Kilgour, R., Laws, M., Taylor, J., Watts, M., Gray, A. (1997a) *Methodology for Speech Data Analysis and a Framework for Adaptive Speech Recognition Using Fuzzy Neural Networks*. ICONIP/ANZIIS/ANNES`97. vol 2, pp. 1055-1060.

10 Keegan, P. (1997) *Kimikupu Hou Maori Lexical Database on the Web: Reflections and Possible Future Directions*. In Proceedings NAMMSAT Conference, October 1997, Massey University, Palmerston North.

11 Laws, M. (1997) *Integrating Text and Speech into Databases, Information Systems and Knowledge Engineering for Human Computer Interaction. The progressive development of an Integrated Bilingual Interface*. In Proceedings NAMMSAT Conference, October 1997, Massey University, Palmerston North.

12 Rheingold, H. (1992) *Virtual Reality*, Mandarin Paperbacks, London.

13 Rob, P. and Williams, T. R. (1995) *Database Design and Applications Development with Microsoft Access 2.0*. McGraw-Hill, San Francisco.

14 Sejnowski, T. J. and Rosenberg, C. (1988) *NetTalk Corpus*, Johns Hopkins University, Cognitive Science Center, Baltimore.

15 Sinclair, S. and Watson, C. (1995) *The Development of the Otago Speech Database*. In Proceedings ANNES 95', University of Otago, Dunedin.

16 Sproat, R. (1996) *Multilingual Text Analysis for Text-to-Speech Synthesis*. 12th European Conference on Artificial Intelligence, Edited by W. Wahlster,

17 Williams, H. W. (1992) *A Dictionary of the Maori Language*, 7th Edition, GP Publications Limited, Wellington.

18 Zue, V. W, Seneff, S., Glass, J. (1990) *Speech Database Development at MIT: TIMIT and Beyond*. Speech Communication 9, pp. 351-356. Elservier Science Publishers, North-Holland.