A CHINESE SPEECH DATA WAREHOUSE

LUK Wing-Pong, Robert and CHENG Chung-Keng
Department of Computing, Hong Kong Polytechnic University
Tel: 2766 5143, FAX: 2774 0842, E-mail: {csrluk,cskcheng}@comp.polyu.edu.hk

ABSTRACT

This paper describes our work in the construction of a data warehouse for speech processing. It is accessible through the internet and requries acess authority. Workers called *contributors* can upload speech files and their corresponding attributes. A relational schema is defined for organizing these attributes into three entities: subject, speaking style and utterance. Before index update, validation ensures the consistency of the attributes since deleting files cannot be done by the contributors. We used a signature indexing scheme because file updates would be frequent. After index update, summary statistics for each entity is updated. For the utterance entity, additional distribution information is collected from the speech files (e.g. coverage of phones). Download is achieved by writing queries. Spelling patterns in queries are specified by regular expressions. At present, only queries that conjoin attributes and disjoin values are allowed. The speech files, their attributes and summary statistics are download in a compressed archive.

1. INTRODUCTION

Whenever there are people, there is human speech sounds. For the same language, speech communities can be scattered through out the world. For example, China has over 50 language communities and the 7 major Hanyu dialects are spread out in the vast territories in China. Native speaker has to be defined by the territories as well as the specific language and dialect. For example, there are differences between Cantonese spoken in Guangzhou and Hong Kong. Thus, A study of speech naturally needs a distributed environment to manage the speech data. In this way, cost (e.g. traveling cost) and duplicate effort can be saved where local institutions collect and analyze speech data of their native speech communities and where they transfer data to a central computer to sharing their resources. Due to limited resources or otherwise, existing corpora are mostly built with one specific dialect or application in mind (e.g. Putonghua from the Hong Kong University [1] and the Institute of Lingusitics, PRC [2], Cantonese from the Chinese University of Hong Kong [3] and Mandarin in Taiwan [4]). Recently, the Linguistic Data Consortium [5] provides a database search facility for their speech data over the World Wide Web. The user can specify speech file attributes to download speech data. Here, we extend this idea further to include the construction of speech corpora and obtaining summary statistics. For the latter, summary statistics are also obtained in data warehouse [6] for business transactions. Hence, we consider our prototype as a data warehouse that would support the collection, search, distribution and summarization of speech data resources.

2. OVERVIEW

Figure 1 shows the architecture of our speech data warehouse. The world wide web is used as the distribution channel of speech data because it is ubiquitous. There are two roles for those interacting with the data warehouse: user and contributor.

The user makes a query to determine the speech corpus it wants to make and download from the data warehouse. After the speech data is downloaded, it is stored in a single file which is exploded into a directory by the speech corpus management (SCM) software. During the speech file extraction, each speech file with its characteristics are registered in SCM so that the user can browse and explore the data. Some simple statistics can also be generated by the SCM, for example duration measurement and balance statistics.

Facilities available to the user are also available to the contributor. In addition, the contributor can send his/her speech data prepared by the data preparation software [7] to the data warehouse and make a request to update the index. Since the speech data file may have virus, only authorized contributors are allowed to add speech data to the data warehouse. The contributor may need to have a local copy of its speech data and therefore a speech corpus submission and management (SCSM) is used. To encourage consistency, the SCM, SCSM and other data preparation software can be downloaded from the data warehouse through the world wide web.

An important issue in corpus construction is whether the corpus is balanced. The data warehouse provides functions to check the balance statistics of the speech corpus and the software for generating speech data of a balanced corpus. An example of balance statistics is the coverage of the different intra- and inter-speaker variables (e.g. diphone, sex, etc.).

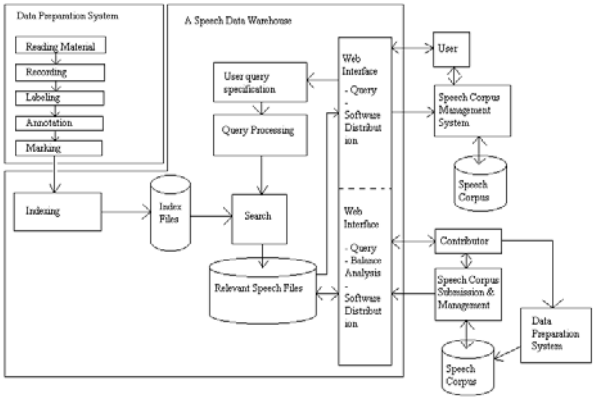

Figure 1: Schematic diagram of our speech data warehouse.

3. DATABASE STRUCTURE

Based on the inter- and intra-speaker variabilities, we have developed a relational database schema for the speech data (Figure 2). The schema is divided into three groups: subject, speaking style and utterance. For the subject entity, the speech file of a particular subject is grouped together. An ID number, which is the primary key, is given to the subject for reference and the name of the subject is not disclosed. For the speaking style entity, different speaking styles are characterized by manner, loudness and speaking rate. An utterance entity is characterized by its transcription, written form, length and the spoken language.

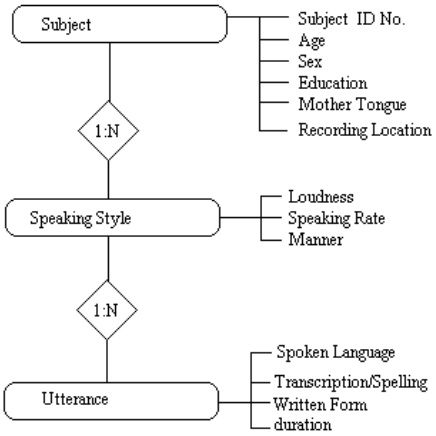

Figure 2: The relational database schema for the speech data

The attributes of the three types of entities enable a wide variety of queries to be formulated. For example, in speech synthesis, the user can download a single speaker speech file by providing the subject ID or by the sex, age and education level of the subject. For research in normalization of accent, the user can obtain speech data where the spoken language is different from the mother tongue of the subject.

4. UPLOAD

4.1 Preparation

Only contributors can submit speech files for sharing. Each contributor is given a unique contributor identification number for security reasons. Before the submission of speech files, the contributor builds a directory structure with the assistance of the SCSM software. Names of the directories are the attribute values of the speech files. For example, if the speech files, *仇.bin* and *史.bin*, are stored in the directory *1/1.3/speed.normal/manner.normal/loudness.whisper/language.Cantonese/* then both speech files are spoken by subject 3 of contributor 1 under normal speed and manner condition but articulated by whispering in Cantonese.

The SCSM software is responsible for the construction of the directory structure in the local computer of the contributor. The SCSM software provides a form for the contributor to fill in the values of the speech characteristics. Since speech articulated in the same or similar conditions is recorded in a set of files, these files should have the same speech characteristics (e.g. speaking rate, manner, spoken language, etc.) and they can be transferred to the same subdirectory according to their speech characteristics.

Once all the speech files are stored in the appropriate directories, they are compressed using the *pkzip* compression utility, recursively for all the subdirectories. The compressed file is uploaded to the data warehouse (Figure 3). Before the upload, the contributor has to enter both his/her identification number and the corresponding password for security.

Figure 3: The upload interface. After the contributor has pressed the *Browse* button, the open file dialog will be appear and the contributor can select the archive to upload.

4.2 Validation

Once the compressed file is uploaded, the data warehouse would decompress the file. Both the speech files and the directory structure would be exploded into the *temp* subdirectory of the corresponding contributor. In this way, clashes during upload between different contributors are avoided.

The data warehouse will *walk through* the entire directory structure and will match the attributes found in the directory structure with the valid attribute names. If no valid attribute names are found, then the attribute extracted from the directory contain errors and this is reported in the validation form (Figure 4). The user can select the desired valid attribute name from the combo box before the master index of the data warehouse is updated.

The subject ID has the form of **<integer>.<Integer>**. The first integer is the contributor ID and the following integer is the ID of the subject given the contributor is known. In this way, clashes between subject ID of different contributors are avoided. Also, for searching, it is easy to search for all subjects of a given contributor.

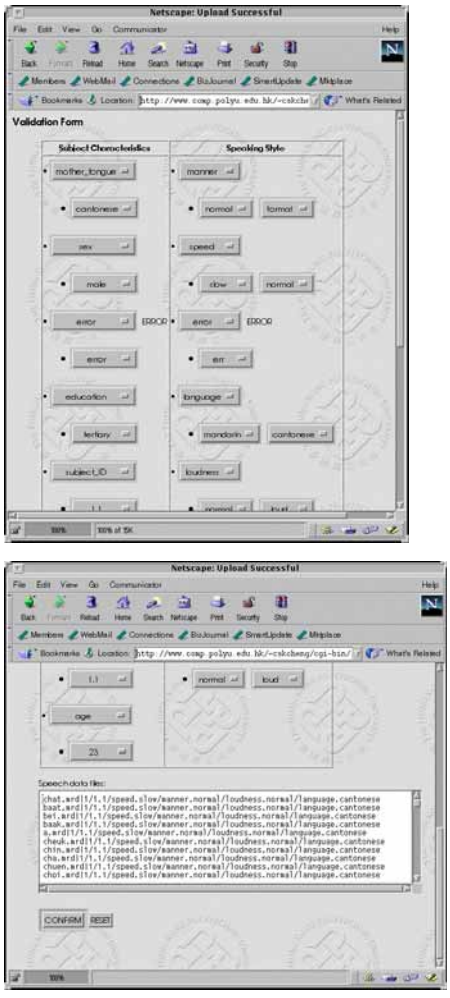

Figure 4: The validation form for an uploaded file. Notice that invalid attribute names are detected with the ERROR label.

4.3 Indexing

When the contributor has confirmed the attribute names are valid, the master index is updated. The master index consists of two tables. One table stores the subject ID and the relevant subject information (e.g. sex, education, etc). The other table stores the speech file names and their corresponding attribute values for the subject ID, speaking style and utterance entity. For indexing, only attributes of the primary key are needed and the second table stores only the subject ID, speaking rate, loudness, manner and spoken language for indexing.

For the second index table, each line has the speech file name and related attribute and related values. In this way, adding new files to the index table is simply appending more data at the end and this index table can be considered as a sequential signature file. For each line, although it is possible to encode specific attribute to a specific position in the line, this encoding becomes invalid when attribute name changes.

Apart from updating the index tables, the speech files are copied from the *temp* directory to the actual directory under the subject ID. The directory structure under *tmp* will also be copied under the subject ID. Since the subject ID is unique and has the contributor ID, there will not be any clashes unless the contributor wants to overwrite or add files to the directory of the subject ID. To avoid contributor overwriting speech files of the subject of the other contributors, the relative subject ID (without the contributor ID preceding it) is automatically appended with the contributor ID, which is obtained when the contributor gives the username (i.e. contributor ID) and the password.

4.4 Summary Statistics

The data warehouse provides some statistics of the speech files for limited monitoring. Logically, we collect statistics for each of the entities: subject, speaking style and utterance characteristics. However, since the statistics are for all the variations of the attributes for the particular entity, the collected statistics would be saved in the directory that contains all the files of that entity. Three files for the three entities will be stored at different levels in the directory structure as shown in Figure 5.

Based on the attribute and values, three types of statistics can be collected. If the attribute is categorical with limited number of categories, then a count can be used for each attribute-value pair. For example, the *sex* attribute has only 2 values. We can count the number of speech files spoken by male and female, separately. If we are counting the number of different values of an attribute, then a count is maintained for the attribute only. An example is the *Subject ID* attribute. Finally, the attribute may contain numerical data, such as the *age* attribute. Here, we maintain and build the distribution with the smallest scale (i.e. 1 unit) so that the distribution with a coarser scale can be re-constructed.

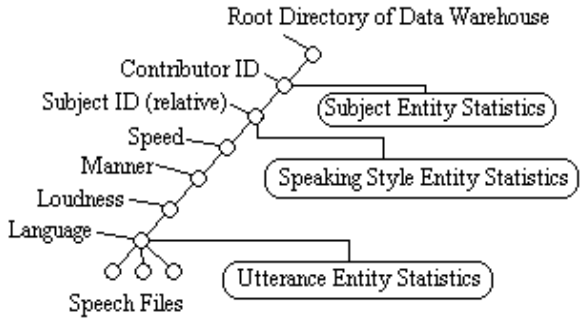

Figure 5: The directory structure of the master index and the corresponding subdirectories that have the statistics for each entity.

For the utterance entity, apart from the statistics for the attributes, additional distribution information is needed to monitor speech corpus construction, as follows:

1) Distribution of speech files with different file formats;

2) Distribution of duration of speech;

3) Distribution of different phones, diphones, initial-final and syllables covered.

The distribution of duration allows us to look at the outliers (i.e. recorded speech that is too short or too long). This may be indicative problems with recording or speech segmentation. Since we are interested in the distribution for the same file format, there is no need to explicitly compute the duration of each speech file. Instead, the distribution of the file storage size would enable us to identifier the duration outliers.

Distribution of the different phonetic units alert us about the coverage and hence the balance of the recorded speech files. Since the concern is for coverage of syllabic and sub-syllabic units, larger unit coverage statistics are not included here. Instead, larger units are segmented into (sub-) syllabic units which distributions would be updated.

The speech files are usually labeled with the written form of the utterance (i.e. Chinese characters). Conversion from Chinese characters to phonetic spelling is carried out for computing the distribution. If the character has multiple spelling depending on context, only the default would be used. To ensure coverage, the contributor has to collect additional information to cover the other multiple spelling of the character.

The speech files may be labeled with the phonetic spelling instead of characters because the Chinese character may not be represented in the computer. This occurs in colloquial text, surnames, place names and dialect-specific characters. Since spelling is specified using ASCII characters, it can be distinguished from Chinese characters using the most significant bit.

Update of the statistics begins from the utterance entity, then speaking style entity and finally the subject entity.

5. DOWNLOAD

To generate a sub-corpus, the user specifies a query (Figure 3). The subject, speaking style and utterance characteristics are listed in a table format. The user clicks the check boxes of the desired attributes. For example, if the user wants to find all female speech, (s)he clicks the check box next to *female* value under the *sex* attribute.

By default, if no value is selected for the attribute, all possible values of that attribute are selected. At present, the query is restricted to boolean queries that conjoin the different attributes and that disjoin the values of the same attribute. For example, if the check box for male and for primary are selected, then the query is *sex=male and education=primary*. If the check boxes for secondary and primary are selected, then the query is *education=(secondary or primary)*.

The user can specify regular expression for phonetic spelling patterns. For example, if the user wants all speech files that starts with a nasal sound, then the following query would be useful: *^(n|ng|m)*. The special symbol ^ indicates that matching begins from the start and | is the disjunction operator. Since the data warehouse is implemented using Perl, the allowed regular expressions are those found for pattern matching in Perl.

Once the query is formulated, searching begins by examining two tables. First, information about the subject IDs is matched between the query and the table containing the ID attributes and values. This will filter the irrelevant subjects. Next, another table containing the rest of the attributes and values is consulted. Here, each speech file, its related attribute and its related values are stored in a single line. Only speech files that have the relevant subject IDs found in the previous search would be examined. Otherwise, the next line is examined. If all the attributes with their corresponding possible values match with the those of the speech file, that speech file would be included as the result. The search ends when all the lines are read.

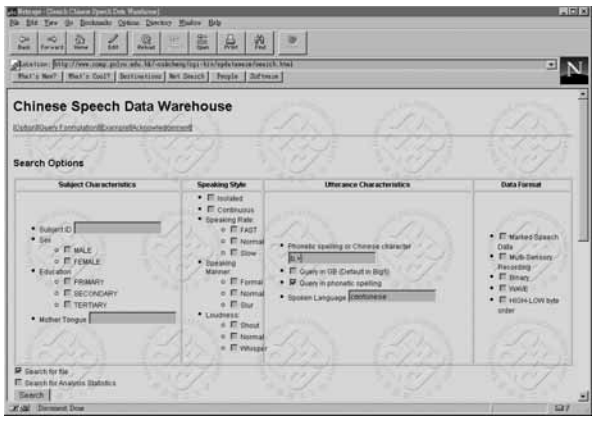

Figure 6: The query interface for the construction of a speech corpus by the user. The user can enter the (partial) phonetic spelling or the Chinese character/phrases for searching the desired speech data. Meta-information (e.g. file format) can also be selected.

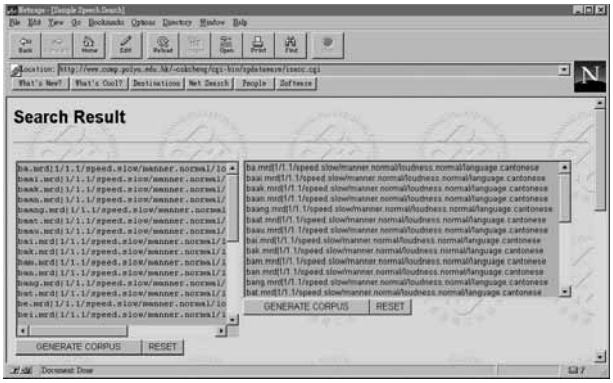

Figure 7: The result of making the query in Figure 2. Individual speech data can be obtained via the hyperlink. Alternatively, all the speech data can be downloaded by pressing the *generate corpus* button. In addition, the user can delete or add items to generate a modified speech corpus.

The matched speech files are collected and put into textboxes for editing or multiline selection. For editing, usually the user would delete a few entries before the speech files are obtained. For multiline selection, the user is expected to select a few speech files for initial examination.

There are two methods to transfer speech files. One method compresses the desired speech files and the directory structure into a zip archive. The compressed file is transferred to the user and the amount of data transfer is reduced. The user can decompress the archive and all the speech files are placed at the corresponding directory locations. For the other method, the user wants all the files to be under the same directory without any directory structure. In this case, the speech files will be added to the archive without the directory information and the list of speech file names and their attribute-values are also added to the archive. Since the speech file names may clash with each other, they will be given a unique integer and the list of speech file names plus their attribute-values is used for searching the desired speech files.

6. FUTURE WORK

For better user-friendliness, we are working on the automatic correction of invalid attributes. This is important because indexing cannot tolerate errors in attribute names and this enforces consistencies between contributors. Also, visualization of the balance statistics would be convenient for the users and the contributors. At present the user visualizes the statistics using Excel or other packages.

A more flexible type of query processing is needed because there are cases where attributes are disjoined together. For example, if a subcorpus is needed for speech sounds produced in a slur manner or at a very fast rate, then two queries are needed. When the number of disjoined attributes is large, writing individual queries becomes impractical.

Queries based on spectral similarity would also be of interest to users because the phonetic spelling may not accurately describe running speech. Even if the speech files are annotated, the user may be interested in spectrally similar speech to examine where phonetic changes occur in running speech.

7. SUMMARY

We have described the general architecture of our speech data warehouse and the related database design. A sequential file is used for indexing because of frequent updates. Apart from speech attributes, statistics of the speech files are also maintained so that query can return the actual files or the summary statistics. At present, our data warehouse is primitive. The validation only examines attribute names without any suggested correction. Later, we aim to provide the best alternative and also validate the attribute values as well. The allowed query is restrictive but adequate for most applications. In the future, we would consider to using inverted-files to support processing of arbitrary boolean queries. Although we do not have speech data covering all the attribute of the database, the availability of such a data warehouse encourage a more coordinated collection and analysis of large language families like Chinese. The data warehouse also assists the worker to maintain a certain degree of balance or coverage of the collected speech data, providing limited monitoring.

ACKNOWLEDGMENT

This work is supported by the (Hong Kong) University Grants Council under CERG PolyU 757/95H.

5. REFERENCES

1. Chan, C. Design considerations of a Putonghua database for speech recognition, *Proceedings of the Conference on Phonetics of the Languages in China*, pp. 13-16, City University of Hong Kong, May, 1998.

2. Zu, Y. (1998) Issues in the software design of reading text for continuous speech database, *Proceedings of the Conference on Phonetics of the Languages in China*, pp. 145-149, City University of Hong Kong, May, 1998.

3. Lo, W.K., K.F. Chow, T. Lee and P.C. Ching (1998) Cantonese database developed at CUHK for speech processing, *Proceedings of the Conference on Phonetics of the Languages in China*, pp. 77-80, City University of Hong Kong, May, 1998.

4. Wang, H.C. Design and implementation of Mandaring Speech Database in Taiwan, *Proceedings of Eurospeech `95*, pp. 875-877, 1995.

5. Linguistic Data Consortium, Access LDC-Online Speech Corpora (hyperlink), http://www.ldc.upenn.edu/online/index.html, 1998.

6. Bischoff, J. *Data Warehouse: practical advice from the experts*, Prentice Hall, 1997.

7. Luk, R.W.P. Speech Annotation by Multi-Sensory ReCording, *COLING-ACL Workshop on Paritally Automated Techniques for Transcribing Naturally Occurring, Continuous Speech*, pp.25-31, Montreal, Canada, August, 1998.