

SpeeData: Multilingual Spoken Data Entry

U. Ackermann², B. Angelini¹, F. Brugnara¹, M. Federico¹,
D. Giuliani¹, R. Gretter¹, G. Lazzari¹, H. Niemann²

¹IRST – Istituto per la Ricerca Scientifica e Tecnologica
I–38050 Povo, Trento, Italy.

²FORWISS – Bayerisches Forschungszentrum für Wissensbasierte Systeme
D–91058 Erlangen, Germany.

ABSTRACT

¹ In this paper we present a multilingual application for speech technology. The SpeeData project aims at building a demonstrator that provides a user-friendly interface for spoken data-entry in two languages: Italian and German. The application domain is the land register of an Italian region in which both languages are officially spoken. The considered data-entry task is particularly challenging as it considers many different types of data - e.g. long texts, numbers, proper names, tables, etc.- and a variety of pronunciations, since dialects are present and users will not always speak in their native language.

1. Introduction

Data-entry can be particularly costly when non electronic information - e.g. contained in documents or pictures - has to be interpreted by a domain expert before being stored into the computer - e.g. medical reporting, diagnostics, cataloging, etc. Recent developments [2] have shown that speech recognition technology has reached a point where it may be used to facilitate such work. The ideal scenario would be to let the computer work like a secretary who types what the expert user dictates, so that she/he does not have to worry about how data are entered and organized into the computer. Further, the user can verify, correct and store the data through a user-friendly interface.

This paper describes the application of multilingual speech recognition to a real life data-entry task. Languages to be recognized are Italian and German. The paper is organized as follows. In Section 1 the application domain is presented, while its multilinguality aspects are treated in Section 2. Section 3 presents the envisaged architecture of the data entry system, Section 4 describes specific multilingual components that will be developed, and finally, Section 5 introduces some evaluation issues of the project.

2. Application Domain

SpeeData will focus on a real application domain: the land register of the Italian bilingual autonomous region Trentino-Alto Adige/Südtirol (RATAA). The land register is an institution that since 1897 has been accumulating information about all rights on real estates. The information are currently maintained in handwritten master books and shall be put into an electronic database. One constraint is that only *relevant* information from the historic master book has to be taken. *Relevance* in this context is defined by laws and represents information that has some impact on today's master book entries. For this work, experts are needed to decide on the importance of the given information.

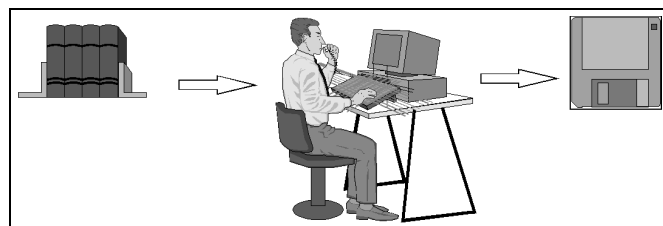


Figure 1: Sketch of the application of the SpeeData project

Today's situation in the land register offices of the region Trentino–Alto Adige/Südtirol consists of two people working on data entry at the same time: the expert extracts information and dictates it to a secretary who types it into the corresponding data base fields. Then, the data base records are printed and corrected by the expert. According to RATAA, a human effort of 285 man years is foreseen for entering data in this way. The introduction of spoken data-entry would be considered a success if labour costs would be reduced by 10 to 20 %.

To enter data to a data base, the information must be structured. A master book entry called *land group* consists of 4 *sheets* according to items like rights, obligations etc. Within the sheets, information can be structured into categories like *name*, *birth date*, *type of right*. Some of the information must be entered in both languages, these fields may be fixed or free text. When using speech as data entry medium, the expert

¹This paper is an abridged version of [1] with modifications.

in general utters a keyword and then the corresponding information. Database fields show a variety of formats:

1. Numbers: they appear mostly when a date, a portion, or an amount of money is mentioned. They may be ordinal, cardinal or fractional numbers.
2. Fixed texts: these texts are limited in their variety. The name of a right or law may be cited, thus, only between 10 and 100 different words are possible at a time.
3. Proper names: they mostly occur when owners of estates are specified, or when geographical locations are indicated (birthplaces, addresses, etc.). Data fields with proper names may require dictionaries of thousands of words.
4. Free texts: this type appears in description of lands or houses. It may also describe some rights among owners and neighbors like *The real estate of X can be crossed by the cattle of Y.*

3. Multilinguality

When working on a multilingual speech recognition system, a good deal of attention must be paid to the languages to be recognized by the system. In this application, a recognition system for Italian and German is built, thus the properties of both languages are of importance.

From the acoustic point of view, the Italian language presents a significantly smaller number of phones with respect to German - e.g. 5 vowels in Italian versus 25 in German. Moreover, recognition experiments on Italian large vocabulary dictation conducted at IRST showed that only minor improvements are achieved with context-dependent (CD) phone models with respect to context-independent (CI) ones. On the contrary, speech recognition experiments conducted at FORWISS shows that significant improvements are achieved by employing CD phone models. This issues impact on the complexity of both the acoustic training and recognition phases, in terms of required data and computations.

From the lexical point of view, both Italian and German are inflected languages. However, German is characterized by a higher variety of flexions and cases [3], a large use of compounding words, and the usage of capital and small letters to specify role of words. All these features heavily reflect on the vocabulary size and on out-of-vocabulary rate, that are in general higher for German (see Table 1).

Newspaper corpus	German	English	Italian
# Words in corpus	31 M	37M	25.7M
# distinct words	650 K	165 K	200K
OOV rate (lexicon 20 K)	10 %	2.5 %	3.7%

Table 1: Comparison of *Wall Street Journal*, *Frankfurter Rundschau*, and *Il Sole 24 Ore* corpora. [6].

For the German language, it can be said, that pronunciation and lexicon strongly depend on the region. Southtyrolean German uses different words and pronunciation rules than standard German. Moreover, the land register experts have either Italian or German mother language and may thus have an accent whenever they enter data in a non-native language. Therefore, the recognition system must not only cope with dialectal variations, but also with a certain amount of accent by the speaker.

4. System Architecture

In this section, the architecture that will be developed for the foreseen system is introduced. This comprises four modules, see Figure 2, namely the *central manager* (CM), the *user interface* (UI), the *data base interface* (DBI) and the *speech recognizer* (SR).

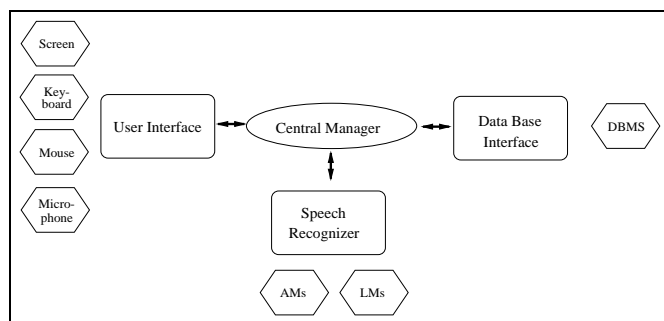


Figure 2: System architecture for the *SpeeData* data entry module.

4.1. Central Manager

The central manager is the core of the architecture. Its tasks are to execute the user's commands, to control and pass information to the other modules, and to maintain the status and the context information of the interaction. During the data-entry, the manager receives speech inputs from the UI and forwards them to the SR together with the specification of the active LM. When the SR has returned a sequence of keywords/field assignments, the manager updates the status and the context information. It also tells the UI module to update the data form on the screen. Moreover, the manager forwards requests to the DBI when data have to be stored or retrieved from the database.

4.2. Database Interface

The data base module is an interface to an external data base management system (DBMS). The formats that the data-entry system and the DBMS use to store entered data need not be the same. The data-entry system only stores a view of the data in the current record, and sends it as a whole to the DBMS for insertion in the data base after user confirmation. During a data-entry session, queries can be generated by the central manager for specific information in the data base.

The role of the data base interface is thus to translate queries or update requests in the DBMS language. In the land register application, the DBMS will be Oracle, since this is the system of choice in the existing keyboard-based system. However, having the DBMS interface isolated to one module, a good portability towards alternative DBMS is provided.

4.3. User Interface

The user interface is the most important unit for the user of the system. Therefore, it must be precisely designed according to the user requirements. The UI manages four devices: screen, microphone, mouse, and keyboard. The last three devices are used for data entry. Using mouse and keyboard, fields can be selected, and data can be entered either by typing or by selecting them from a menu (if possible). Using speech, the user can perform the same actions. Moreover, in many forms the user will be allowed to fill in more fields with a single utterance and without specifying the single fields. The three input modalities can complement each other: for example, the mouse can be used to select a certain field, the information may be given via speech.

The screen fulfills several tasks: feedback, context, guidance. First of all, it provides feedback for the user about the entered data. This comprises both the possibility of error correction and knowledge about which data have been entered during the current session. Furthermore, the screen shows stored information about the current land register entry as a result of previous input or in relation to another stored land register entry. With the screen, a medium for guidance is provided. The user finds at any time a set of proposals which information might be entered next. Additionally, before finishing an entry, the user will see if all necessary information is given or if something is missing.

4.4. Speech Recognizer

The task of the speech recognition module is to process speech events. This module will be treated in detail in the next section. Information between the speech recognition module and the central manager flows in two directions. A stream of speech events reaches the speech recognizer, then the information is transformed into a word sequence. Additionally, it is decided, in which field(s) the data are stored. If there is a keyword, it is transformed into information concerning the field. If no keyword is entered, the system has to choose by the content of the speech itself and by context-specific rules.

5. Multilingual Components

SpeeData aims at integrating a speech recognizer derived from the IRST dictation system [5]. The core of the system is a time-synchronous, beam-search Viterbi decoder where language models are represented by means of finite state networks.

Up to now, this system has already been successfully employed in dictation systems related to different domains. However, peculiar features of the data-entry task, namely multilinguality and language model switching, will require the system to be adapted. Moreover, multilinguality will reflect on the way user-adaptation of the system will be carried out.

5.1. Acoustic Modeling

This issue aims at providing a multi-lingual speech recognizer system, where the different languages are integrated as smoothly and transparently as possible. Language differences come into play both at the acoustic and the linguistic level. For a phonetic point of view, the first step of a unified approach is to adopt the SAMPA phonetic alphabet, which covers the lexicon of both languages with a unique set of phonetic units. The translation of SAMPA units in actual recognition units (HMMs) is however not straightforward. It is known, for example, that the impact of CD phone models on word accuracy is language dependent, and in particular it differs for the two considered languages (see Section 3). It is thus conceivable that, though keeping the common paradigm of the SAMPA base units, different levels of detail will be chosen when designing the acoustic models for the two languages, trying anyway to avoid a too big impact on the overall system complexity.

5.2. Language Model Switching

As for language model switching, this is a feature needed because utterances corresponding to different data-fields will have different linguistic constraints, i.e. different language models. This is in contrast with a dictation system, when the LM is usually fixed. To cope with this problem, the system will have to be able to efficiently switch the active LM according to the data-entry system state. This could be done by simply loading at system startup a set of predefined LMs, one for every possible interaction state, and then selecting one among these alternatives every time the speech recognizer is invoked. After a preliminary analysis, however, this approach has been abandoned, because it would require the compilation of many different LMs which could be only slightly different from each other. Instead, it has been decided that there will be a core set of "primitive" LMs, through combination of which the actual LMs to apply at different states will be built on-the-fly by a specific module. This approach, though requiring more work for the adaptation of the recognition system, will allow much greater flexibility and resources economy.

5.3. Speaker Adaptation

An other feature of a data-entry task is that the typical user is not occasional. There will be people that will use the system for a long period, so it will be not necessary to rely on speaker-independent (SI) units only. While SI units should provide a satisfactory baseline for a new user, a speaker-

adaptation module will be available, allowing to better focus the acoustic modeling on the particular speaker. Both *batch* and *incremental* adaptation techniques will be considered. In batch adaptation, users will be asked to utter a few tens of sentences in both languages, during an enrollment session. Phonetically reach sentences will be designed that provide a balanced phonetic content for both languages. Then, a *maximum a posteriori* estimation technique will be applied to adapt the parameters of the acoustic units. Incremental adaptation will instead occur during system usage and will not require any effort by the user.

5.4. Language Model Adaptation

Adaptation will concern language models as well. In this case the adaptation will take place with respect to different sites. There will be two kinds of LM adaptation. First, since the LMs will be class-based, site adaptation could consist in filling a general class (e.g. proper name) with site dependent data (e.g. the set of the most frequent proper names of a specific district). Secondly, LM probabilities could be adjusted as well, when enough sample data become available at the target site.

6. Evaluation and Measurements

6.1. Performance Evaluation

Speedata will require refinements of existing speech recognition technology owned by the partners. Performance tests will be carried out to monitor both internal improvements and the state-of-the-art in the field. For this reason, domain independent test suites will also be considered. Evaluations will focus on all the single components of the speech recognizer: acoustic models, language models, search algorithm, language adaptation algorithm, speaker adaptation algorithm, etc.

Two main quality characteristics will be addressed: accuracy and efficiency. In fact, the ultimate goal of each component will be to improve accuracy of the recognizer without trading off efficiency, i.e. consumption of computational resources (memory and CPU time).

6.2. Utility Evaluation

The aim of the system is to allow data-entry of land register books by speech. The demonstrator must allow a single user to carry out this job and to store the same information as with a traditional keyboard based system. Functionality of the system will be first analysed through empirical methods and finally, i.e. during the assessment-on-site phase, through objective measurements. In particular, the *task-coverage rate* will be measured by computing the percentage of data-entry tasks that can be managed with the system.

6.3. Usability Evaluation

Usability refers to all those aspects of the demonstrator that involve human-computer interaction. Usability must be considered carefully, as the most important features of the system lie in its speech recognition based user interface. Usability attributes for user interfaces can be found in [7]. For the Speedata project learnability, efficiency, memorability, error rate, and satisfaction are the attributes that are analysed. The analysis can happen by means of heuristic evaluation, thinking aloud tests, performance measures, and questionnaires.

7. Acknowledgements

This work is supported by the European Commission, Telematics Application Programme, project reference number LE 1999. Other partners in the Speedata project are Informatica Trentina (Italy), Regione Autonoma di Trentino Alto-Adige/Südtirol (Italy) and Bundesministerium für Justiz (Austria). The views and conclusions contained in this document are those of the authors.

8. REFERENCES

1. U. Ackermann, F. Brugnara, M. Federico, and H. Niemann. Application of speech technology in the multilingual Speedata project. In *Proc. CRIM-FORWISS Workshop*, Montreal, 1996.
2. B. Angelini, G. Antoniol, F. Brugnara, M. Cettolo, M. Federico, R. Fiutem, and G. Lazzari. Radiological Reporting By Speech Recognition: The A.Re.S System. In *Proc. ICSLP*, vol. 3, pp 1267–1270, Yokohama, 1994.
3. H. Bußmann. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart, 2nd edition, 1990.
4. W. Eckert, T. Kuhn, H. Niemann, S. Rieck, A. Schueer, and E.G. Schukat-Talamazzini. A spoken dialogue system for German intercity train timetable inquiries. In *Proc. EUROSPEECH*, vol. 2, pp 1871–1874, Berlin, 1993.
5. M. Federico, M. Cettolo, F. Brugnara and G. Antoniol. Language Modelling for Efficient Beam-Search. *Computer Speech & Language*, 1995.
6. L. Lamel and R. De Mori. Speech recognition of European languages. In *Proc. of the IEEE ASR Workshop*, Snowbird, 1995.
7. J. Nielsen. *Usability Engineering*. Academic Press, Boston, 1993.
8. E.G. Schukat-Talamazzini and H. Niemann. ISADORA — A Speech Modelling Network Based on Hidden Markov Models. *Computer Speech & Language*, 1993.