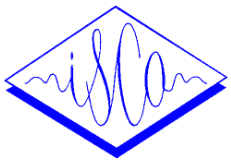

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

RADIOLOGICAL REPORTING BY SPEECH RECOGNITION: THE A.Re.S. SYSTEM

Bianca Angelini, Giuliano Antoniol, Fabio Brugnara, Mauro Cettolo,
Marcello Federico, Roberto Fiutem and Gianni Lazzari

IRST-Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo (Trento), Italy

ABSTRACT

Radiological reporting has already been identified as a field in which voice technologies can prove to be very useful. Recent progress in automatic speech recognition and in hardware and software technology makes it possible to build large-vocabulary, continuous speech, speaker-independent, real-time systems.

In this paper a dictation system for radiology reporting, the A.Re.S. system, is presented. A.Re.S. is a "software only" system which runs in real-time on an HP 715 workstation. It relies on an asynchronous and multi-process architecture in which speech decoding is performed by processes in pipeline.

System requirements and architecture will be described, together with the results of a preliminary evaluation based on three months of on-site testing.

I. INTRODUCTION

Recent progress in Automatic Speech Recognition (ASR) and in hardware and software technology makes it possible to build large-vocabulary, real-time, speaker-independent systems. Medical document generation presents features which render it a good application field for ASR [5, 6], since the employed language is typically standardized and the vocabulary size is manageable. In the particular field of radiological reporting, for example, computer dictation allows a radiologist to dictate naturally a report while looking at the X-ray photograph.

In this paper an application of speaker independent, continuous speech, and real-time ASR for radiology is presented. The system, named A.Re.S. (Automatic Reporting by Speech), covers a vocabulary of 6,000 words, related to the emergency examinations domain. A prototype of A.Re.S. has been installed at S. Chiara Hospital[1] and is daily used by physicians.

II. THE A.Re.S. PROJECT

2.1 Radiological Reporting

A radiological report is the typed document that describes and summarizes the physician's observation deducted from an X-ray photograph.

In the daily routine, reports are usually recorded by physicians on tape cassettes and then transcribed by secretaries. In emergency medicine, where report generation time is critical, a secretary instead types the report under physician's dictation.

[1] S. Chiara Hospital is a regional structure that serves a large community of about 100,000 people; its Radiological Department performs more than 150,000 examinations a year.

In both radiology and emergency medicine, report generation can be automated using an ASR system, which supplies a draft text transcription that needs only to be revised by the secretary or by the physician her/himself.

2.2 Project Overview

The A.Re.S. project has been developed at IRST in collaboration with the Radiological Department of S.Chiara Hospital, Trento. It started in 1990 with a feasibility study to identify requirements and propose technologies [3]. The first prototype [4], delivered in May '92, allowed the dictation of reports related to chest examinations (2,500 words of vocabulary), with a small pause between words, and in speaker dependent modality. It ran almost in real-time on a HP 720 workstation under Unix, and used a DSP board to perform feature extraction. The current A.Re.S. prototype overcomes the limitation of isolated word dictation by employing a speaker independent and continuous speech recognition engine.

2.3 System Description

Following the distinction between routine and emergency reporting, A.Re.S. provides two operation modalities: *batch* and *interactive*.

In batch modality, A.Re.S. follows a tape recorder metaphor. A digital voice recorder is simulated which resembles a normal tape recorder with the exception that it does not record pauses between words or sentences. The report, dictated through a microphone, is stored in a logical cassette and is transferred in a waiting queue of cassettes, which represents the set of recordings confirmed by the user and to be processed by the speech recognizer. Besides dictating, the physician has the option of associating a spoken memo to a report - e.g. to identify the patient.

When recognition of a report is completed, the cassette on the top of the waiting queue is eliminated and put in the queue of ready reports. These reports are now available for verification and correction. They can be loaded into an editing environment, where they can be verified, by listening to the corresponding signal, edited and printed.

The interactive modality differs from the batch one in that the signal samples are directly sent to the recognizer, while the physician is dictating, and the generated text is immediately available in the editing environment.

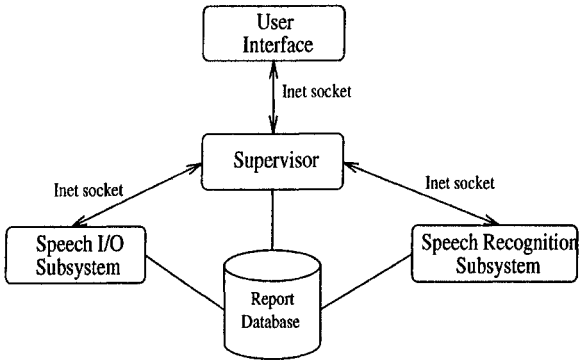

Figure 1: *System Architecture.*

III. SYSTEM ARCHITECTURE

The A.Re.S. system is based on a multi-process, master-slave architecture, consisting of a central supervisor that activates and coordinates the remaining subsystems and also maintains the global state of the system (see Figure 1).

The peripheral subsystems are:

- the User Interface, which provides the user with a window-based command and editing environment;
- the Speech I/O Subsystem, which supplies the capabilities for signal acquisition, recording and synthesis;
- the Speech Recognition Subsystem, which performs transcription of dictated reports.

Each subsystem is connected with the supervisor through a communication channel, implemented with Unix Internet sockets. With this communication scheme, processes can run on different hosts connected by a local area network.

3.1 Supervisor

The Supervisor module maintains the global system state and consistency. Further, it also acts as a communication server for the other subsystems.

During the startup phase, the Supervisor activates the peripheral subsystems, creates a communication channel with each one, and initializes each subsystem. Then it waits for messages from the subsystems.

Message reception is asynchronous and requests are stored on a FIFO queue, for successive processing. The Supervisor allows the execution of user requests, according to its internal state. It can be thought of as a finite state automaton, which implements the Supervisor's behavior, with two macro states, namely Dictation and Editing.

3.2 User Interface

The graphical user interface (shown in Figure 2) is organized in two functional layers: *presentation* and *dialog control.*

The presentation layer manages creation and handling of graphic objects. It is implemented with the Motif toolkit, and consists of four objects: the digital voice recorder, the queue of the recorded cassettes, the window of the completed reports, and the word processor. This layer also transforms the input device events (keyboard and mouse) into logical events that are appropriate for the subsequent layer.

The dialog control layer handles events coming from the user and events coming from the Supervisor

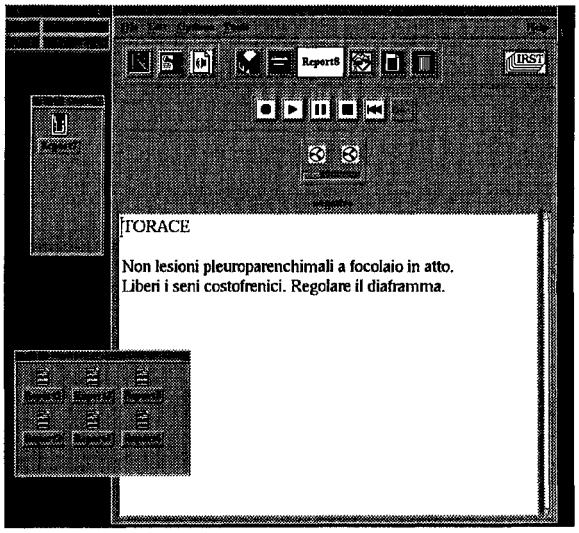

Figure 2: *Graphical User Interface. It consists of three windows: the main window, with the digital voice recorder and the word processor, the window of the cassette queue (top left) and the window of completed reports (bottom left).*

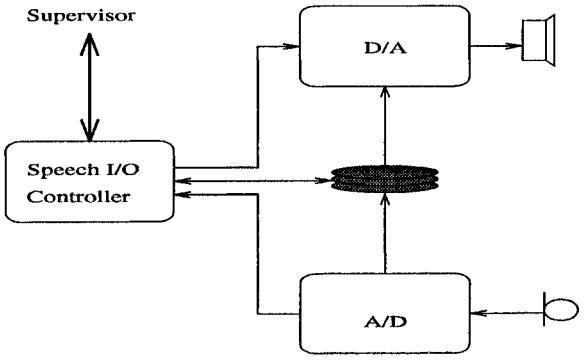

Figure 3: *Speech I/O subsystem.*

through a communication channel. It is implemented as a finite state automaton that evolves through different states according to the received events, executes *presentation actions* towards the user, and sends messages to the Supervisor.

3.3 Speech I/O Subsystem

Pure software ASR systems may have problems in handling Analog to Digital (AD) and Digital to Analog (DA) conversions. In order to make sure that data to be collected is not lost, a layered architecture (see Figure 3) has been designed in which a controller manages two independent modules for DA and AD conversions.

The layered approach allows the same AD/DA controller to be used in both system modalities: in the interactive one, where the recognizer is immediately fed with the incoming signal, and in batch modality, where the signal is stored on disk.

Having two independent modules for AD/DA conversions also makes software portability and reusability easier. In fact, these two modules are the only ones whose implementation is platform dependent.

Even if modern workstations allow high fidelity

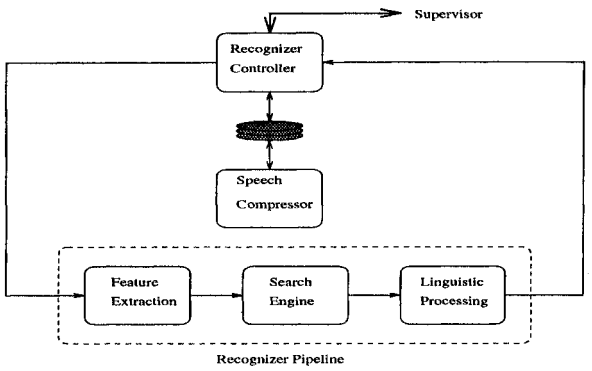

Figure 4: *Speech Recognition subsystem.*

sound recording, a 16kHz sampling rate is used as it is adequate for speech recognition applications.

3.4 Speech Recognition Subsystem

A layered structure has also been adopted for the Speech Recognition Subsystem. A Recognizer Controller manages the data flow between the Supervisor and either the disk storage or the Recognition Module. The Recognition Module is a pipeline of three processes:

- Feature extraction, computing the acoustic parameters needed by the search engine;
- Search Engine, which decodes the signal representation into a sequence of words;
- Linguistic Processing, which puts the Search Engine output into an appropriate form for the user.

IV. ASR IN A.Re.S.

4.1 Feature Extraction

The signal processing front-end provides the recognizer with a 27-dimensional vector every 10ms, consisting of 8 MEL scaled cepstral coefficients, the log-energy, and their first and second time derivatives. The acoustic parameter vector is scaled so as to ensure that all its elements have comparable ranges.

4.2 Acoustic Modeling

The phonetic transcription of words uses 37 context independent units. Unit HMMs have simple left-to-right topologies of three or four states, depending on the average length of corresponding units. Distributions are gaussian mixtures with a variable number of components, resulting from a training process which initializes all mixtures with 24 components, and then prunes less used gaussians. The final configuration used in the experiments reported includes a total of 1858 gaussians grouped in 198 mixtures. Unit models were trained on the APASCI acoustic database collected at IRST [1]. The training set comprises about 2,000 phonetically rich sentences which are completely unrelated to the radiology domain.

4.3 Beam-search

The recognition system employs a one-pass Viterbi beam-search. This approach has shown to be an effective way of dealing with huge search spaces [7] without losing accuracy. The objective of Viterbi search, which can be seen as an application of the Dynamic Programming principle, is to find a globally optimal sequence of weighted arcs on a network. The beam-search technique consists in neglecting states whose accumulated score is lower than the best one minus a given threshold. The procedure of selectively discarding unlikely paths has the implicit effect of localizing in time the influence of the acoustic observations on the search.

A characteristic of a consistent implementation of the beam-search strategy is that the decoding time mainly depends on the number of expanded active arcs and not on the size of the whole search space.

Our implementation of the decoding algorithm takes into account the fact that the units network can be huge, hence, in spite of the network being statically represented, the memory used in intermediate computations is allocated on demand, with a simple caching strategy.

In evaluating gaussian mixtures, instead of summing all the terms, the approximation of taking the most likely term is done, since this allows a time gain without affecting accuracy. Moreover, caching of distribution values is performed, ensuring that every distribution is computed at most once on every frame, even if the same model appears on many arcs.

4.4 Language Modeling

A vocabulary was selected from a collection of about 50,000 computer written reports, with the aid of a physician. In order to estimate a bigram based Language Model (LM), one half of the reports was corrected for spelling errors, while only in-vocabulary word sequences were extracted from the remaining part.

Since reports were dictated and transcribed by different physicians and typists, respectively, some further preprocessing was necessary in order to cope with the many different abbreviations, acronyms and jargon which physicians and typists use - e.g. "sn", "sin." "sx" are all abbreviations of "sinistra" (*left*). Some very frequent synonyms (e.g. the example above) have been clustered into word classes on which bigrams are estimated, in order to have a more general LM and to allow people to use their favorite expressions. Class based bigrams have been also introduced for numeric expressions which exploit a decomposition of numbers into units, tens, hundreds, etc. Word sequences corresponding to numbers are then collapsed into a numeric expression by a post-processor which operates on the recognizer output. Bigrams were estimated with an interpolation formula (see [2]) which allows a useful LM representation into a finite state network.

4.5 Search Space Organization

The bigram LM has been represented by a tree-organized finite state network. A tree representation of the phoneme-transcription of words allows the sharing of their common beginning phonemes and is suited to the beam-search strategy better than other kinds of search-space arrangements [2]. Integration of the bigram LM into the tree-based organization of the lexicon required the explicit (tree) representation of successor words which were seen during LM training. The network so obtained was successively reduced by an optimization algorithm. The size of the resulting network is well within the memory capabilities of a workstation.

V. Performance Evaluation

5.1 Laboratory Tests

During development of the 6,000 word prototype installed at the hospital, some testing material was col-

lected in order to assess system performance. On 330 reports, uttered by one male speaker, recorded in a noise insulated room, the prototype achieved a word accuracy of 95.6%.

In a second phase another more complete acquisition started. A continuos speech database of radiological reports has been collected for the purpose of carrying on more reliable tests. The database contains 759 reports read by 4 different speakers (3 male and 1 female physicians) for a total of almost 5 hours of continuos speech. Only one of the subjects was acquainted with the system.

Each speaker completed her/his acquisition in three separate sessions. Reports to be read were chosen so that the largest number of radiological terms was covered.

The information available for each dictated report is its orthographical transcription, which exactly reflects what was uttered, and the reference text, where abbreviations or acronyms appear. For both transcriptions a segmentation by sentences is also available.

Recordings were performed in a noise insulated room. Speech was acquired at 48kHz, with 16 bit accuracy, by means of a Digital Audio Tape-Corder Sony TCD-D10PRO and a super cardioid microphone Sennheiser MKH 416-T. Digital signals were then filtered and downsampled to 16kHz to reflect the actual working conditions.

Recent experiments were performed on the 4-speaker database with an extended vocabulary of 10,000 words and a larger set of unit models [2]. By setting search engine parameters in order to have real time response, a word accuracy of 93.0% was achieved.

5.2 On-site Tests

The 6,000 word prototype has been tested on-site over a period of three months. During this time the system was used by 5 physicians and all input signals, recognizer outputs and revised texts were saved. A preliminary and rough evaluation of this material provided a word accuracy of 86%. In fact, more refined measurements are difficult as in several reports entire phrases were changed by the physicians regardless of what they had dictated, hence reference texts should be corrected by listening to all the recordings.

VI. DISCUSSION

Working on a real-world application presented aspects which are not normally considered in research activity. Design specifications were directly suggested by the end user, namely a physician who collaborated to this project from its begin. In fact, designing a machine to be embedded in a well established working environment needed to carefully consider how people really work, in order to let the system be well accepted. As a matter of fact, people tend to be quite conservative for what concerns their working habits. Further, performance from the user's point of view does not only mean high recognition accuracy and fast response, but also well formatted printed reports.

A crucial point of this work was the availability of text samples by which model the language of the application. About 2 millions of words were made available by S. Chiara Hospital. Data was produced by typists writing under dictation. This condition, which seemed at a first glance optimal, hided indeed several problems. We realized that the data was affected by many kinds of distortions - i.e. spelling errors, grammar errors, different ways of abbreviating words, etc.. This required lots of efforts in order to extract a consistent data set to train the language model.

The assessment of the system on-field turned out to be a difficult task. The user acceptance can be considered satisfactory on the basis of the physicians impressions, but it could not be given a quantitative evaluation. Moreover, the problems mentioned in the previous section made evaluation of recognition accuracy imprecise.

VII. CONCLUSIONS

A dictation system for radiological reporting has been presented together with some evaluation experiments. Both the functionalities and the software architecture of the system have been described.

The system architecture allows execution of the subsystems on different hosts. Hence, it can be configured as a single speech recognition server connected with many low cost clients devoted to speech entry.

Further, an overview of the speech recognition techniques employed in the system has been given. Experiments in laboratory and on site showed the effectiveness of the developed ASR technologies in this application field.

Finally, laboratory results of a new version of the system, working with 10,000 words, have been given.

ACKNOWLEDGEMENT

The authors wish to thank Dr. Enzo Moser for his support in defining the system requirements and in analyzing the training data.

References

[1] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo. Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus. In *ICSLP*, Yokohama, Japan, 1994.

[2] G. Antoniol, F. Brugnara, M. Cettolo, and M. Federico. Language model estimations and representations for real-time continuous speech recognition. In *ICSLP*, Yokohama, Japan, 1994.

[3] G. Antoniol, F. Brugnara, F. Dalla Palma, G. Lazzari, and E. Moser. A.RE.S.: an interface for automatic reporting by speech. In *Proceedings of the European Conference on Speech Communication and Technology*, Genova, Italy, 1991.

[4] G. Antoniol, R. Fiutem, R. Flor, and G. Lazzari. Radiological reporting based on voice recognition. In Leonard J. Bass, Juri Gornostaev, and Claus Unger, editors, *Human-computer interaction : third International conference, EWHCI: selected papers*. Lecture Notes on Computer Science, Springer-Verlag, Berlin, Germany, 1993.

[5] H. Cerf-Danon, S. DeGennaro, M. Ferretti, J.Gonzalez, and E. Keppel. Tangora - a large vocabulary speech recognition system for five languages. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 215–218, Genova, Italy, September 1991.

[6] R. Joseph. Large vocabulary voice-to-text systems for medical reporting. *Speech Technology*, 4(4):49–51, 1989.

[7] H. Ney, D. Mergel, A. Noll, and A. Paesler. Data driven search organization for continuous speech recognition. *IEEE Transactions on Signal Processing*, February 1992.